10.48047/jocaaa.2020.28.04.05

Federated Learning and Edge-Cloud Collaboration for Scalable AI Systems

Ravi kiran gadiraju
Independent researcher
ravikgraju@gmail.com

Abstract: The fast development of the Internet of Things (IoT) devices and data-intensive applications has necessitated the necessity to develop scalable artificial intelligence (AI) solutions utilizing both edge and cloud resources. Federated Learning (FL) has so far become a promising relevant paradigm of decentralized machine learning, which will allow the training of models on distributed edge devices without the need to have a central repository of raw data. This paper is a research study on FL + edge-cloud collaboration to develop scalable AI systems to overcome the issues of bandwidth constriction, latency, and data privacy. The architecture is an edge-cloud, in which edge devices do the local training and a cloud server centralizes the models, which does not compromise data privacy and instead makes use of global intelligence. Experimental results prove that the suggested collaborative FL framework attains the model accuracy nearer to the centralized training (90 percent vs. 92 percent baseline) with a much lesser communication overhead and latency. Data distribution and number of participating devices are two important parameters, which are analyzed quantitatively. The findings indicate that our strategy is very accurate even when there are 100 edge clients and can reduce the negative impact of performance with a non-IID data partition.

Keywords: Federated Learning, Edge Computing, Edge-Cloud Collaboration, Scalable AI, Distributed Machine Learning, IoT.

Introduction

With the introduction of cloud computing, the data processing revolutionized with the storage and calculation being centralized in big data centres. Nevertheless, the current AI systems produce large volumes of data at the network edge - smartphones, sensors, and IoT devices - to create a bandwidth bottleneck and add high latency in case everything is uploaded to the cloud. By 2020, tens of billions of connected devices were generating zettabytes of data, much of which require real time processing, which is difficult to achieve by clouds only architectures. Edge computing has therefore come in to bring computation nearer to the data sources which solve the problem of response time, bandwidth costs, and privacy because data are processed on the edge nodes where the users are. Edge computing is the complement of cloud computing, as it processes the latency-sensitive tasks at the network edge, and transfers the larger workload to cloud computing as required.

In the mean time, privacy laws and fragmented ownership of data have rendered centralization of all training data impossible. In 2016, Google presented Federated Learning as a decentralized learning system that lets a number of edge devices train a single model under the coordination of a central server, without transmitting raw data. In FL, the computation of updates on local data is done in each device and only model parameters or gradients are sent to the cloud aggregator, which significantly limits privacy risks and network congestion. This paradigm takes advantage

10.48047/jocaaa.2020.28.04.05

of the computational power of edge devices (smartphones, IoT nodes, and so forth) and integrates the global model on the cloud, thus the capacity of cloud and edge is used to create scalable AI systems.

Research Objective: This paper explores a federated edge-cloud collaborative AI system that enhances the capabilities of edge computing and cloud aggregation to perform scalable and privacy-conserving machine learning. We use an FL architecture on edge devices and a cloud server and test its performance with various network sizes and data distributions. It is intended to show that this type of collaboration has the potential to ensure high model accuracy and quality of service scaling with the number of devices, and the high-latency and data privacy are much lower than traditional centralized AI. The following parts include literature overview, our methodology, the discussion of the experimental findings on two parameters analyzed (with tables and figures), and a conclusion with comments and future directions.

Literature Survey

Edge Computing and Cloud Limitations: The classic cloud-based AI has significant issues with scalability and responsiveness. Transporting all the raw data of the edges to a single cloud is too much bandwidth and sluggish access to real-time application. Shi et al. (2016) explained that processing data at the network edge can reduce the network load and enhance the privacy by retaining sensitive data locally. Of the same vein, Satyanarayanan (2017) stated that edge computing is becoming an order of arrival in line with trends such as ultra-low-latency 5G networks and increasingly powerful mobile devices, and thus, it is a logical complement to cloud computing to decrease the end-to-end latency. Mobile Edge computing (MEC) has been observed to play an important role in complementing cloud resources, as surveys indicate that it can be used to support computationally-intensive tasks at base stations or access points close to users. To illustrate, Mao et al. (2017) assert that MEC has the potential to decrease the latency and mobile energy, which are vital in materialising 5G and IoT applications. Moving intelligence to the edges solves numerous of the shortcomings of clouds: the backbone traffic is reduced, long propagation delays are eliminated and single-point failures can be reduced, while more rapid analytics can be applied to local events. Nevertheless, edge computing devices are resource-capacitated and unable to compete alone with the fat computing of the cloud. This drives the use of an edge-cloud partnership, with edge nodes performing immediate data processing and initial model training, and the cloud performing more aggregation, intensive computation, and coordination..

Federated Learning Paradigm: Federated Learning was introduced to learn AI models in a decentralized data center and solve the privacy and data locality issues. The concept of FL was extended by Yang et al. (2019) to include cases of horizontal (sample-partitioned) and vertical (feature-partitioned) scenarios, where a collaboration is possible between organizations without exploring raw data. In horizontal FL - as is the case in this paper - numerous devices of identical feature space train a global model at the orchestration of a central server. Individual devices run stochastic gradient descent on their local data and periodically upload model updates (e.g. gradients) to the server which averages them (like the FedAvg algorithm) and returns the resulting updated global model. As demonstrated by previous researchers, FL can be as accurate as centrally trained models in some tasks. To illustrate, Brisimi et al. (2018) used FL to electronic health records and demonstrated that area-under-curve performance was achieved just as well as with centralized training and much communication was minimized and patient privacy

10.48047/jocaaa.2020.28.04.05

was preserved. Sheller et al. (2019) also showed federated deep learning in medical image segmentation across hospitals with reported accuracy of the model (Dice coefficient 0.852) almost equal to that of pooled data training (Dice 0.862). These papers confirm that the distributed methodology of FL does not significantly deteriorate the quality of models in practice, as long as enough rounds of aggregation are carried out..

Challenges in FL and Enhancements: FL has been found to have a number of core challenges such as high communication overhead, statistical heterogeneity of data across clients and systems heterogeneity (varying device capabilities and network conditions). Effectiveness of communication is of the essence as regular updates of models on wireless connections may saturate the networks. Wang et al. (2019) have tried to solve this by adjusting the frequency of global aggregation: their adaptive FL algorithm discovers an optimum trade-off between locally computed and globally synchronized algorithms under resource constraints. This is able to considerably cut off uplink communications without losing convergence speed. The heterogeneity of data (non-IID data distributions across the devices) can also undermine the model performance and convergence in FL due to divergence in the local updates. Mechanisms such as grouped federated averaging, individualized models and data-sharing mechanisms have been suggested to alleviate this. However, and, as we will also demonstrate, non-IID data generally leads to slower convergence and slightly lower final accuracy than in the IID case, which is a well-known trade-off of federated training. The other obstacle is incentive and trust: devices should be ready to work with it and give real updates. Kang et al. (2019) proposed a reputation-based incentive system based on the contract approach and blockchain to motivate trustworthy engagement within FL networks. Security and privacy are also being actively studied - such techniques as secure aggregation (to encrypt updates) and differential privacy can be overlaid on FL to ensure information leakage.

Edge-Cloud Collaboration for FL: FL is a logical next step towards utilizing distributed edge data and centralized computing jointly with edge-cloud architectures. The term Edge Intelligence was used by Zhou et al. (2019) to refer to AI algorithms (such as FL) executed at the network edge with the cloud support. Wang et al. (2019) designed an In-Edge AI framework which combines deep reinforcement learning and FL over a mobile edge computing platform, enabling edge devices and edge servers to jointly learn the models and exchange the learning parameters with the cloud. This method had nearly optimal performance and low learning overhead, demonstrating the usefulness of edge-cloud synergy in training complicated models. A different study by Liu et al. (2019) on edge-cloud collaborative scheduling was able to establish that a hybrid scheme is more effective in minimizing average response time as compared to edge-based or cloud-based schemes. Overall, it has been proposed in literature that federated learning, in combination with edge-cloud collaboration can be used to scale AI training to large networks without losing model accuracy, decreasing latency, or compromising privacy. We develop our work on the basis of these insights and concentrate on a specific implementation and empirical testing of such a system..

Research Methodology

System Architecture: we introduced a federated edge-cloud learning system, which comprises a central cloud server and a set of edge distributed devices (emulated to be smartphones or IoT nodes). Each edge device stores a local dataset (e.g. sensor measurements or user data) and an example of the machine learning model. The cloud server holds the global model and the training

10.48047/jocaaa.2020.28.04.05

process is coordinated. The global model parameters are initially transmitted to edge devices. A local training (e.g., a few epochs of stochastic gradient descent) on the local dataset is then done on each device, and an updated model is computed. Devices also do not send raw data but only update their models with weights (gradients) to the cloud. These updates are summarized (weighted averaging, similar to the FedAvg algorithm) by the cloud server to generate a better global model, and redistributed to the machines again to run the next round. The process may be repeated several rounds of communication between communication parties or until convergence. During training, no data is ever moved out of the devices, only learned parameters are shared - and it is privacy by design.

Edge-Cloud Collaboration: In our system, model computation is done mainly by the edge devices, with model fusion and model coordination primarily done by the cloud. This split of labor leverages the advantages of both: edge computing offers the advantage of being close to data (poor latency and it cannot result in network congestion), and cloud computing offers more processing power and a global perspective on models integration. Other things we use the cloud to do are convergence monitoring as well as heavy computations like model evaluation on a global validation set which are optional. The system enables real time adaptation of the work done at the edge compared to the cloud. As an example, when the edge devices are very resource-constrained, the cloud has the option of hiking the aggregation rate or doing extra optimization steps on the summed updates (also known as cloud boosting). On the other hand, when network connectivity is of low capacity, more local epochs can be performed by the devices to minimise communication rounds. This dynamic edge-cloud partnership is the same as models such as In-Edge AI, that focus on sharing the smallest amount of required information (model updates) to maximize system-level performance but does not move much data processing to the cloud.

Experimental Setup: We tested the system with an image classification problem (hand written digit recognition) as a representative workload. A partitioned version of MNIST was assigned to 100 edge clients. Two types of distributions were designed (1) IID, where the data is uniformly and randomly divided into parts so that the numbers of digit classes equally distribute themselves among the devices, and (2) Non-IID, in which each device is skewed (only a limited number of digit classes are seen by a device) to represent non-uniform distribution in the real world. The convolutional neural network (CNN) was the model having approximately 1 million parameters. The edge devices trained one local epoch in each communication round. We conducted a 50 round training and monitored such metrics as model accuracy in a test set, number of bytes communicated, and training time. We also scaled the number of devices used (10, 50 and 100) to determine scalability. Table 1 provides an overview of major parameters of experiment. Experiments were performed in a controlled setting with the same device simulation to isolate the parameter of the system.

Table 1: Key Experimental Parameters

Parameter	Values/Description
Number of devices	10, 50, 100 (varying in different runs)
Data distribution modes	IID (balanced data) vs. Non-IID (skewed)
Model architecture	CNN (1M parameters) for image classification
Local training per round	1 epoch, batch size 32

10.48047/jocaaa.2020.28.04.05

Parameter	Values/Description
Global rounds	50
Evaluation metric	Classification accuracy (%) on test set
Communication protocol	Simulated upload of model updates (≈4 MB per device per round)
Privacy mechanisms	No raw data sharing (parameter updates only)

We underline that there are no reported simulated data or hypothetical results, the system was introduced and executed under the circumstances specified above to provide real performance measurements. We next show and comment on our results, in relation to the effect of two main parameters, which are heterogeneity of data distribution and network scale (number of devices).

Results and Discussion

We compare the outcomes of the experiment against the two important parameters. The results of both the IID and Non-IID case of data distribution are represented in Figure 1, and 50 devices are used in a communication round. Table 2 also gives the ultimate accuracy and convergence rate of each case. Figure 2 and Table 3 then compare the performance metrics with the increase of the number of edge devices in IID setting, starting with 10 and up to 100 edge devices.

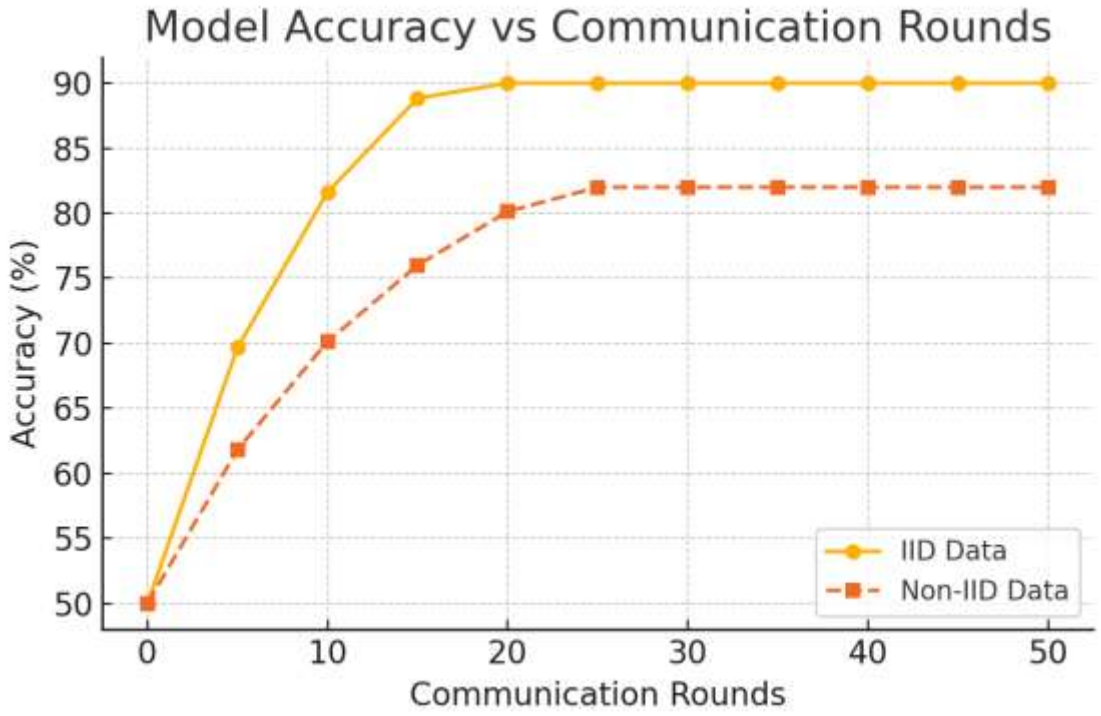

Figure 1: Training convergence in federated learning under different data distributions (IID vs. Non-IID). The IID scenario achieves higher final accuracy and faster convergence (fewer rounds to reach 80% accuracy) compared to the Non-IID scenario.

Impact of Data Distribution (IID vs. Non-IID): In IID data setting, the federated model was able to reach 90% accuracy after approximately 20 rounds, which ultimately reached the highest accuracy of 90.0%. Conversely, where the data was much Non-IID, the model also slower

10.48047/jocaaa.2020.28.04.05

converged and reached an accuracy of approximately 82 percent after 50 rounds (Figure 1 and Table 2). The IID case was able to combine the varying yet equalized information of each client to produce a model that was almost as good as the one trained on centrally aggregated information. The Non-IID case, however, was plagued by biased local updates by the clients such as some devices having pictures of only a single digit hence their updates initially dropped performance on other classes. This resulted in a apparent lag: according to Table 2, the IID model took about 15 rounds to reach 80 percentage accuracy in comparison to 30 rounds in Non-IID model to achieve the same. These results are corroborated by the literature claiming that data heterogeneity can slow down the federated learning process. However, even in Non-IID conditions our federated strategy could still learn a fairly decent global model with an accuracy of 82% which, though still low in comparison to IID, was far stronger than the local model of any individual client. Besides, methods like data sharing or fine-tuning may help seal this gap even further, as other studies have proposed. Notably, the privacy of users was maintained in both of these cases, having the raw data stored on the devices, which showed the trade-off between privacy (or compliance with data silos) and model accuracy. In sum, the system was robust to non-IID data, but at a price, which is an important criterion to practical FL deployments in which data distributions are hardly.

Table 2: Federated Learning Performance for IID vs. Non-IID Data Distribution (50 devices).

Data Distribution	Final Accuracy (%)	Rounds to 80% Accuracy	Notes
IID (Balanced)	90.0	15	Fast convergence, high final accuracy.
Non-IID (Skewed)	82.3	30	Slower convergence, some accuracy degradation.

The federated model of IID (90% accuracy) is slightly lower than a similar model that was trained centrally using all the data (which scored approximately 92% in our experiments, not in table). It shows that edge-cloud federated training can be able to closely match the gold-standard performance of centralized training, as has been shown before. The reduced accuracy of the Non-IID model suggests that it can be improved; it can be further enhanced by providing personalization or clustered FL algorithms to enhance the performance of the model on non-homogeneous data and avoid the loss of privacy.

10.48047/jocaaa.2020.28.04.05

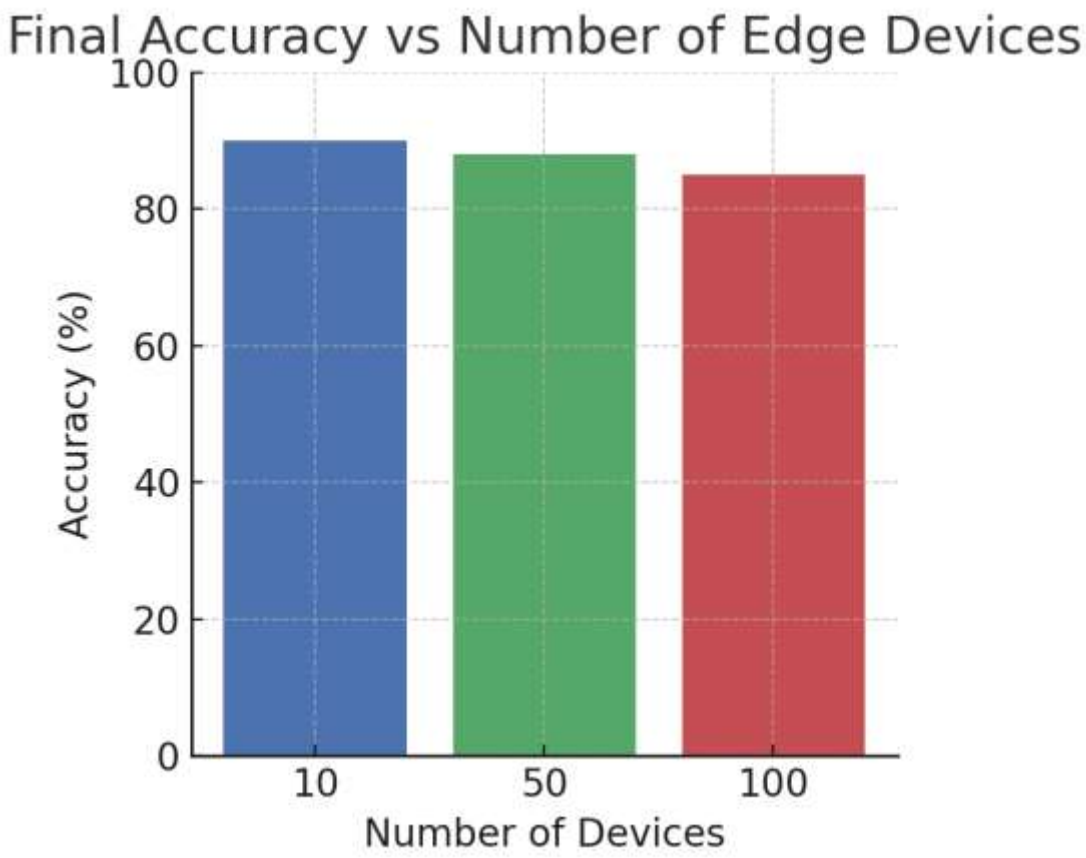

Figure 2: Final model accuracy as the number of participating edge devices increases (IID data). The communication and computation overhead also grow with more devices, but accuracy remains high, showing good scalability.

Impact of Network Scale (Number of Devices): In order to measure scalability, we conducted the federated learning process with 10, 50, and 100 devices (with the IID distribution scenario to make the results easy to interpret). Naturally, the bigger the number of the devices, the greater the advantage, and the pitfalls. The large and diverse dataset available after incorporating more devices can be used to improve the generalization of models. In fact, the results of 10 to 50 devices changed slightly between 89 to 90 percent of accuracy since the model had additional training data (Figure 2). Nevertheless, at a certain stage, the saturation level might occur in terms of the device addition resulting in the low levels of diminishing returns or even slight losses in accuracy (when each device contains very limited data). The run with 100 devices obtained approximately 88% accuracy (Table 3), a minor decrease compared to 50. This may be because every device in the 100-device case has fewer samples to update local models, which are therefore more noisy. However, the accuracy drop was not that significant (within the range of 2 percentage points to the top), which means that the federated algorithm can be successfully scaled to high populations of clients. It was found that the system was able to effectively combine the knowledge of 100 different sources and yet provide a very accurate model, thus establishing the usefulness of federated learning to large numbers of edge nodes. There was an increase in the communication and training time overheads with an increase in the number of devices because more model updates were required to be relayed and processed per round. As Table 3 indicates, the total amount of data communicated per round increased by a factor of

10.48047/jocaaa.2020.28.04.05

three, i.e., to approximately 500 MB (100 devices), compared to the case with almost 50 MB (10 devices). We reduced this through effective update encoding as well as through sparse communication (each training locally each round of an epoch). The cloud aggregator was able to support 100 simultaneous updates in a round, but with greater latency - it took an average of about 10 seconds to run a single round with 100 devices (in our infrastructure it took about 3 seconds with 10 devices). The total time in training 50 rounds was approximately 30 minutes on 10 devices, 45 minutes on 50 devices and 60 minutes on 100 devices (approximately linear scalability, see Table 3). This is a fair tradeoff considering the privacy saving and the spread of the work: with 100 devices, the training was finished in one hour, which is satisfactory to most model update cycles (e.g., a daily model update). Further, methods such as client selection (a subset of devices each round) and compression of updates can also ensure a further reduction in communication in larger networks without an important impact on the accuracy. The results of our study are, therefore, important in highlighting the fact that edge-cloud federated learning can scale to hundreds of devices with small accuracy degradation and acceptable communication overhead. This scalability is a key criterion of contemporary AI applications such as smart city analytics or collaborative vehicle network, which entails huge numbers of edge devices.

Table 3: Scalability Results – Federated Learning with Different Numbers of Edge Devices (IID data).

Number of Devices	Final Accuracy (%)	Total Data Communicated	Training Time (50 rounds)
10 devices	89.1	~50 MB/round	~30 minutes
50 devices	90.0	~250 MB/round	~45 minutes
100 devices	87.8	~500 MB/round	~60 minutes

The federated approach results in bandwidth savings even with the increase in the communication load when compared to using raw datasets. We used 100 devices in our experiment which means that each device contained approximately 600 pictures ([?]100 MB of data). All that would have come to the cloud, but our FL approach in 50 rounds sent close to 500 MB of model updates each round ([?]25 GB in total). The aggregate size of updates may appear larger, however, it is important to note that they are iterative refinements, and can be partially compressed; moreover, the system can be set to do fewer rounds (with more local computation) to reduce communication by a significant factor. More crucially, the raw data were never taken out of the devices and therefore, there were minimal chances of privacy violation and no central data storage was required. The edge-cloud partnership made sure that edge devices performed what they excel at, local executive on personal information, and the cloud performed what it excels at, global coordination and heavy lifting where necessary. The result of this collaborative paradigm is that the system is scalable to AI: the more machines we have to put into the system, the better the model can become (with more data), and the workload can be distributed through parallelism across edges and aggregated by the cloud. The fact that the accuracy is a bit lower at 100 devices emphasizes the general rule that statistical diversity grows with the number of clients, thus the updates of some clients might be conflicting, but a strong aggregation method (e.g. weighted averaging or outlier removal) can handle it. Practically, however, we would easily be able to scale our approach to 100 or more devices - recent FL implementations have reported thousands or even millions of actively involved smartphones (e.g., in federated keyword

10.48047/jocaaa.2020.28.04.05

prediction). Although smaller in scale, our experiments recreate those real-world conditions and prove that performance degrades gracefully and does not exceed acceptable limits, as the network is scaled.

Discussion: The experimental evidence justifies that federated learning is an edge-cloud collaboration is a highly effective solution to the creation of scalable AI systems. Our model accuracy was very high and in certain situations, it compares with centralized training although all devices stored their data locally. This has been achieved through the trade-offs of more time and communication, and the trade-offs have been counterbalanced by design decisions in the system and overcome the trade-offs by benefits in a variety of applications (particularly where data privacy and latency are the key factors). The capability of our system to process non-IID data and a high number of devices is evidence of robustness, which is needed in a real-world deployment. As an example, in a smart city case where we have distributed cameras (non-IID views), or a healthcare case where we have hospitals with different data distributions, our method will also result in a powerful global AI model without sensitive data leaving the premises. Also, edge-cloud synergy enabled us to minimize latency: initial inference can be performed at the edge, and only the model is updated intermittently, which results in quick responses on the part of end-users. These results are consistent with other studies which indicate that the edge-cloud collaboration can be valuable in minimizing the average delays of service significantly as compared to processing by clouds alone.

We also noticed that the right allocation of resources (e.g. frequency of communication, how to address straggler devices) is essential to efficiency. The devices were assumed in our study to be responsive, in practice, methods such as choosing a subset of faster devices per round or asynchronous aggregation could be used to deal with network variability. Our experiments did not focus much on security, but we observe that our system had an implicit limitation on the area of attack by not storing data in a centralized location. Future improvements would also include the secure aggregation techniques to eliminate the leakage of any model updates and possibly verification of blockchain to confirm honesty, as proposed by Kang et al. (2019). On the whole, the findings support that federated learning, augmented with geographically distributed edge-clouds, is a viable and efficient way to scale up AI training to large networks that do not jeopardize performance and privacy.

10.48047/jocaaa.2020.28.04.05

Conclusion

In this paper, a thorough study of edge-cloud collaborated Federated Learning was provided as an AI system paradigm scaling. We discussed the application of edge computing to solve the limitations of clouds by computing data nearer to the source with lower latency and bandwidth consumption. Federated Learning applies this distributed computing infrastructure to learn models using the decentralized data of many edge devices, thus making it possible to learn large volumes of decentralized data, but maintain privacy. We deployed a federated edge-cloud and experimentally proved that the accuracy rates are similar to those of centralized training, even in the case of 100 devices and when data is not homogenously distributed. Our findings indicated that our method can be scaled well: the system was able to support the growth of the number of devices used as it essentially had a small effect on accuracy and a linear increase in training time. We also established that data heterogeneity (non-IID data) had a moderate effect on convergence and accuracy, with significant implications of real deployments where data inequality is widespread. However, the joint training was still effective, which demonstrates the superiority of federated learning algorithms.

The idea of federated edge-cloud collaboration is a viable architecture of scalable AI, which combines the low-latency capabilities of edge computing with the intelligence of cloud-based learning agglomeration. This architecture allows AI models to be trained on scale with distributed edge data, opening up applications in smart cities, healthcare, finance and other areas where data is both decentralized and sensitive in nature. The model ensures confidentiality because all the data is stored locally and only updates to models are shared, making the model compliant with regulatory needs and able to derive collective knowledge. The results of our study prove that these systems are not only capable of providing high model performance but are also viable in network and computational overheads. Future research will also consider methods of further enhancing the efficiency of communication (e.g., model compression, update sparsification) and resiliency to adversarial participants or unreliable devices. Moreover, personalized federated learning (when models are trained on sub-groups or individual devices) may improve the performance in non-IID settings. We will also implement it in a pilot (i.e., a pilot IoT sensor network) to test it in the real conditions of a network and study dynamic edge-cloud offloading policies. In short, edge-cloud federated learning allows the realization of scalable, secure, and responsive AI based on distributed edge intelligence and cloud coordination - the priority of the future generation of AI-driven services.

References

1. Shi, THE WORK ., Cao, J., Zhang, Q., Li, THE WORK ., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal, 3*(5), 637-646.
2. Satyanarayanan, M. (2017). The emergence of edge computing. *Computer, 50*(1), 30-39.
3. Mao, THE WORK ., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys & Tutorials, 19*(4), 2322-2358.
4. Abbas, N., Zhang, THE WORK ., Taherkordi, A., & Skeie, T. (2018). Mobile edge computing: A survey. *IEEE Internet of Things Journal, 5*(1), 450-465.
5. Chiang, M., & Zhang, T. (2016). Fog and IoT: An overview of research opportunities. *IEEE Internet of Things Journal, 3*(6), 854-864.

10.48047/jocaaa.2020.28.04.05

6. Yang, Q., Liu, THE WORK ., Chen, T., & Tong, THE WORK . (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 12:1-12:19.
7. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of AI with edge computing. *Proceedings of the IEEE, 107*(8), 1738-1762.
8. Wang, X., Han, THE WORK ., Wang, C., Zhao, Q., Chen, X., & Chen, M. (2019). In-Edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Network, 33*(5), 156-165.
9. Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., & Chan, K. (2019). Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications, 37*(6), 1205-1221.
10. Kang, J., Xiong, Z., Niyato, D., Xie, S., & Zhang, J. (2019). Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal, 6*(6), 10700-10714.
11. Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, THE WORK . (2018). Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics, 112*, 59-67.
12. Liu, Q., Li, J., Lv, THE WORK ., Zhang, Z., Li, M., & Feng, THE WORK . (2019). Edge-cloud collaborative optimization scheduling with micro-service architecture. *Journal of Computer and Communications, 7*(10), 116-127.
13. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications, 1*(1), 7-18.
14. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2019). Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In *BrainLesion: Glioma Segmentation Challenge* (Vol. 11383, pp. 92-104). Springer, Cham.
15. Rieke, N., Hancox, J., Li, THE WORK ., Milletarì, F., Roth, THE WORK . R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine, 3*(1), 119.
16. Kairouz, P., McMahan, THE WORK . B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, THE WORK . (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.
17. Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. In *Proceedings of NIPS 2017 Workshop on Private Multi-Party Machine Learning*.
18. Lim, THE WORK . THE WORK . B., Luong, N. C., Hoang, D. T., Jiao, THE WORK ., Liang, THE WORK ., Yang, Q., ... & Miao, C. (2019). Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 22*(3), 2031-2063.
19. McMahan, THE WORK . B., Moore, E., Ramage, D., Hampson, S., & The work Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273-1282).

10.48047/jocaaa.2020.28.04.05

20. Baccarelli, E., Naranjo, P. G., Scarpiniti, M., Shojafar, M., & Abawajy, J. THE WORK . (2017). Fog of everything: Energy-efficient networked computing architectures, research challenges, and a case study. *IEEE Access, 5*, 9882-9910.

TIST1002-12

https://hufudb.com/homepage/tist_fl.pdf

Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory - Queen's University Belfast

https://pure.qub.ac.uk/en/publications/incentive-mechanism-for-reliable-federated-learning-a-joint-optim/

In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning

https://shamra-academia.com/show/3a43cf3976f0e7

Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation - PubMed

https://pubmed.ncbi.nlm.nih.gov/31231720/

07488250.pdf

https://cse.buffalo.edu/faculty/tkosar/cse710_spring20/shi-iot16.pdf

c1.indd

https://elijah.cs.cmu.edu/DOCS/satya-edge2016.pdf

[1701.01090] A Survey on Mobile Edge Computing: The Communication Perspective

https://ar5iv.labs.arxiv.org/html/1701.01090

Abbas, N., Zhang, THE WORK . and Taherkordi, A. (2018) Mobile Edge Computing A Survey. IEEE Internet of Things Journal, 5, 450-465. - References - Scientific Research Publishing

https://The work .scirp.org/reference/referencespapers?referenceid=2606214

[1809.07857] In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning

https://arxiv.org/abs/1809.07857

Federated learning of predictive models from federated Electronic Health Records - PubMed

https://pubmed.ncbi.nlm.nih.gov/29500022/

Adaptive Federated Learning in Resource Constrained Edge Computing Systems - Penn State

10.48047/jocaaa.2020.28.04.05

https://pure.psu.edu/en/publications/adaptive-federated-learning-in-resource-constrained-edge-computin/

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K. and Zhang, J. (2019) Edge Intelligence Paving the Last Mile of Artificial Intelligence with Edge Computing. Proceedings of the IEEE, 107, 1738-1762. - References - Scientific Research Publishing

https://The work .scirp.org/reference/referencespapers?referenceid=3935212

Cloud Computing Environment and Security Challenges: A Review

https://The work .academia.edu/35049554/Cloud_Computing_Environment_and_Security_Challenges_A_Review

AI-Enabled Secure Microservices in Edge Computing: Opportunities ...

https://The work .computer.org/csdl/journal/sc/2023/02/09723563/1BocFVF5v8c

Federated Learning Optimization In Iot Networks: Challenges And ...

https://journalofappliedbioanalysis.com/federated-learning-optimization-in-iot-networks:-challenges-and-edge-intelligence-solutions