netidee

STIPENDIEN

Communication and Energy Efficient Edge Artificial Intelligence Framework for Internet of Things

Endbericht | Call 18 | Stipendium ID 6801

Inhalt

1 Introduction....3

2 General....3

3 Results....4

4 Planned Follow-up Activities....5

5 Suggestions for Continuation by Third Parties....5

6 Bibliography....6

1 Introduction

The research is focused on developing communication and energy efficient Edge-AI framework to overcome the limitations of centralized learning to pave the way for futuristic applications enabled by widespread adoption of smart but resource-constrained IoT devices. The literature review conducted during this work led to the discovery of the most critical issue namely the inability to cater to the needs geographically distributed clients faced by the current proposed solutions which consequently led to the development of a novel federated learning framework termed pHFedKD (personalized Hierarchical Federated Learning via Knowledge Distillations) to address such limitations. This document encompasses and presents the general information regarding the research conducted during this project, outcome of the project along with the planned future research activities, and last but not the least the set of guidelines for the interested parties including academics and the likes for extending or utilizing our work.

2 General

The proliferation of IoT sensors have given rise to a new set of futuristic applications such smart cities, smart environmental and precision medicine. These constellations of smart devices scattered across the globe generate massive amounts of data and hold potential to addressing some of the most pressing issues of the contemporary times such as climate change. Although the advancements in IoT and AI present new opportunities at the same time they pose significant challenges to traditional centralized computing due to the share volumes of the data being generated at the edge of network. It puts a massive strain on the network to transfer this much amount of data. Additionally, transporting data from remote places to a centralized server for data analysis raises service concerns about privacy and latency. This mandates developing efficient Edge-AI systems for resource-constrained devices to mitigate these challenges.

Federated learning (FL) allows distributed clients to collaboratively train a model without requiring them to share their private data. Each client is supposed to keep their data private instead only shares the weights of its locally trained model. After a few rounds of local training, clients share their updates with the global parameter server which aggregates these updates from the clients to generate the global model. FL works well in homogeneous settings in which clients' data are sampled from the same distribution. However, its performance deteriorates significantly in the face of data heterogeneity or more formally known as non-IID problem. As a direct consequence of this, FL struggles to be as effective in large environments with a large number of geographically dispersed

clients. Moreover, frequent communication between the clients and single central server could incur huge communication costs.

The core objective of this work was to develop a framework leveraging the entire spectrum of computing continuum which addresses the aforementioned challenges. The research question that we set out to find an answer for was “can we leverage the client heterogeneity to our advantage instead of merely treating it as a problem? “. The way we do is first by quantifying the heterogeneity (diversity) among clients and later utilizing that for segregating homogeneous (similar) clients into clusters. Subsequently, the complex task of learning a single global model is divided into smaller tasks by assigning each cluster to an edge server and learning cluster-specific edge models. Hence, we develop a framework, which leverages client heterogeneity to learn individual client, cluster, and global model simultaneously.

3 Results

We introduce a hierarchical federated learning framework called pHFedKD which addresses client heterogeneity, concept drift, and scalability in FL for decentralized environments with diverse client data distributions. The proposed framework combines dynamic clustering, multi-level hierarchical aggregation, and multi-teacher knowledge distillation to optimize performance of the client, cluster, and global models.

Our key contributions made to the domain through this work are listed below:

- We introduce federated geospatial clustering (FGC) based on a novel client affinity (CA) metric to quantify data and client similarity. This dynamic clustering method adapts to spatio-temporal client mobility and concept drift by reassigning clients during training.
- We introduce a hierarchical federated learning framework which spans three computing continuum levels; client, edge, and cloud. At level 0, clients train local models using their private data. At level 1, edge server aggregate local models to generate cluster-specific models. At level 2, cloud server aggregate cluster models to yield the global model.
- We introduce multi-teacher knowledge distillation to enable inter-cluster knowledge sharing and enhance global and edge models. This ensures that clusters are able to retain their unique characteristics while simultaneously benefiting from the shared global knowledge.

We conduct through evaluations of our proposed framework using real-world CityScapes dataset for semantic segmentation in both static and dynamic settings and compare it to

the state-of-the-art baselines such as FedAvg, FedProx, and HierFavg. pHFedKD outperforms baselines under both static and dynamic conditions, achieving substantial gains in model accuracy and communication efficiency. More specifically, it achieves up to 19% improvement in client-level accuracy and 24.7% in edge-level accuracy. Furthermore, it reduces communications costs up to 50% transmitting only lightweight logit vectors instead of the entire weight vectors.

4 Planned Follow-up Activities

The finished work is currently submitted for publication and under review at the IEEE International Conference on Parallel and Distributed Processing Symposium (IEEE IPDPS). The developed framework was evaluated on an autonomous vehicle urban scene understanding use case; however, it can easily be extended to other domains. Validation of its effectiveness on additional use cases is ongoing. Currently, we are adapting the framework for an intelligent farming application focused on tracking animals in remote alpine regions.

In parallel, the preparation of the dissertation is underway. The next planned steps include:

- Submitting the complete draft for internal review and incorporating feedback.
- Finalizing the dissertation and submitting it formally for evaluation.
- Coordinating with the examination committee to schedule the defense.
- Preparing for the oral defense through presentation rehearsals and potential Q&A simulations.

These steps are aimed at completing the dissertation process and achieving a successful defense in the coming months.

5 Suggestions for Continuation by Third Parties

The proposed framework can be applied in diverse real-world applications such as autonomous systems, smart cities, and precision agriculture, where heterogeneity, mobility, and data privacy and real-time requirements are critical. Furthermore, pHFedKD offers potential for interoperability studies to explore integration with existing federated learning frameworks and edge computing infrastructures.

Third parties could investigate incorporating advanced clustering methods, such as meta learning-driven adaptive clustering, to further optimize dynamic grouping. Additionally, integrating pHFedKD with privacy-preserving mechanisms, such as differential privacy or secure multi-party computation, could expand its applicability in domains with stringent

regulatory requirements, like healthcare or finance. Finally, pHFedKD can be studied alongside resource optimization techniques to explore trade-offs between computational cost and personalization, particularly in energy-constrained environments like IoT devices.

6 Bibliography

[A+22] Ehsan Amid et al. Robust federated learning through representation matching and adaptive hyper-parameters. In *International Conference on Learning Representations*, 2022.

[ABKS99] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28:49–60, 1999.

[AOG20] Mohammad Hosseini Abad, Emre Ozfatura, and Deniz Gunduz. Hierarchical federated learning across heterogeneous cellular networks. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8866–8870, 2020.

[AVTP+17] Ferran Adelantado, Xavier Vilajosana, Pere Tuset-Peiro, Borja Martinez, Joan Melia-Segui, and Thomas Watteyne. Understanding the limits of lorawan. *IEEE Communications magazine*, 55(9):34–40, 2017.

[BDC+18] Taoufik Bouguera, Jean-François Diouris, Jean-Jacques Chaillout, Randa Jaouadi, and Guillaume Andrieux. Energy consumption model for sensor nodes based on lora and lorawan. *Sensors*, 18(7):2104, 2018.

[BEG+19] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, et al. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019.

[BFA20] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–9. IEEE, 2020.

[C+21] Liam Collins et al. Exploiting shared representations for personalized federated learning. *International Conference on Machine Learning*, pages 2089–2099, 2021.

[Cis20] Cisco. Cisco annual internet report (2018–2023) white paper. Technical report, Cisco Systems, 2020.

[CWL24] Keyan Cao, Jian Weng, and Kuan-Ching Li. Edge computing and cloud computing for internet of things: A review. *Future Internet*, 16(3):94, 2024.

[DDT21] Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021.

[Deb01] Kalyanmoy Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley Interscience Series in Systems and Optimization. John Wiley & Sons, Chichester, UK, 2001.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[DKM20] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

[DTN+21] Canh T Dinh, Nguyen H Tran, Minh NH Nguyen, Choong Seon Hong, Wei Bao, Albert Zomaya, and Vincent Gramoli. Federated learning over wireless networks: Convergence analysis and resource allocation. *IEEE/ACM Transactions on Networking*, 29(1):398–409, 2021.

[ESBT21] S Mehdi Emadian, F Oyku Sefiloglu, Isil Akmehmet Balcioglu, and Ulas Tezel. Identification of core micropollutants of ergene river and their categorization based on spatiotemporal distribution. *Science of the Total Environment*, 758:143656, 2021.

[F+24] Muhammad Shoaib Farooq et al. Role of iot technology in agriculture: A systematic literature review. *Electronics*, 13(2):357, 2024.

[FMO20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

[GCYR20] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[Gun18] Nyoman Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018.

[GYMT21] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[H+21] Samuel Horvath et al. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34:12876–12889, 2021.

[H+23] Seyyedali Hosseinalipour et al. Federated learning over wireless networks: Convergence analysis and resource allocation. *IEEE/ACM Transactions on Networking*, 31(1):398–413, 2023.

[HAMS21] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[JOK+18] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

[K+60] Rudolph Emil Kalman et al. A new approach to linear filtering and prediction problems [j]. *Journal of basic Engineering*, 82(1):35–45, 1960.

[KKM+20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143, 2020.

[KNH14] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. online: http://www. cs. toronto. edu/kriz/cifar. html, 55(5), 2014.

[KSG08] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

[KSZ08] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452, 2008.

[KW03] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[KXN+19] Jiawen Kang, Zehui Xiong, Dusit Niyato, Yuze Zou, Yang Zhang, and Mohsen Guizani. 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, 2019.

[L+24a] Yang Li et al. Topology-aware federated learning in edge computing: A comprehensive survey. *ACM Computing Surveys*, 56(8):1–35, 2024.

[L+24b] Shiqiang Liu et al. Expanding the cloud-to-edge continuum to the iot in serverless federated learning. *Future Generation Computer Systems*, 157:421–432, 2024.

[LHM+18] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

[LHY+20] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *Proceedings of the International Conference on Learning Representations*, 2020.

[LKSJ20] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, pages 2351–2363, 2020.

[LSTS20] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[LSZ+20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[LZSL20] Lumin Liu, Jun Zhang, SH Song, and Khaled B Letaief. Client-edge-cloud hierarchical federated learning. In *ICC 2020-2020 IEEE international conference on communications (ICC)*, pages 1–6. IEEE, 2020.

[LZW20] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

[M+23] Jon Mills et al. Multi-task federated learning for personalised deep neural networks in edge computing. *Pervasive and Mobile Computing*, 93:101801, 2023.

[Mea08] Donella H. Meadows. *Thinking in Systems: A Primer*. Chelsea Green Publishing, White River Junction, VT, 2008.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[MYZ+17] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B Letaief. A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys & Tutorials*, 19(4):2322–2358, 2017.

[NY19] Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE International Conference on Communications*, pages 1–7. IEEE, 2019.

[PKP+19] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[RKS17] Usman Raza, Parag Kulkarni, and Mahesh Sooriyabandara. Low power wide area networks: An overview. *IEEE Communications Surveys & Tutorials*, 19(2):855–873, 2017.

[SCST17] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4424–4434, 2017.

[SCZ+16] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016.

[SJW18] Gal Shulkind, Stefanie Jegelka, and Gregory W Wornell. Sensor array design through submodular optimization. *IEEE Transactions on Information Theory*, 65(1):664–675, 2018.

[SMS21] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3710–3722, 2021.

[SYS+20] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, 2020.

[TD20] Javid Taheri and Shuiguang Deng. *Edge Computing: Models, Technologies and Applications*. The Institution of Engineering and Technology (IET), 2020.

[TDK24] Hoa Tran-Dang and Dong-Seong Kim. Distributed learning in the iot–edge–cloud continuum. *Machine Learning and Knowledge Extraction*, 6(1):283–315, 2024.

[TZ20] Yi Tan and Limao Zhang. Computational methodologies for optimal sensor placement in structural health monitoring: A review. *Structural Health Monitoring*, 19(4):1287–1308, 2020.

[VPS21] Shanu Verma, Millie Pant, and Vaclav Snasel. A comprehensive review on nsga-ii for multi-objective combinatorial optimization problems. *IEEE access*, 9:57757–57791, 2021.

[WC20] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[WHL+20] Xiaofei Wang, Yiwen Han, Victor CM Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(2):869–904, 2020.

[WTS+19] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

[WYS+20] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations (ICLR)*, 2020.

[YA08] Mohamed Younis and Kemal Akkaya. Strategies and techniques for node placement in wireless sensor networks: A survey. *Ad Hoc Networks*, 6(4):621–655, 2008.

[YJL+21] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, pages 12073–12086. PMLR, 2021.

[YST20] Kwonjoon Yu, Ruslan Salakhutdinov, and Josh Tenenbaum. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9339–9348, 2020.

[Z+24a] Chen Zhang et al. Enabling federated learning across the computing continuum: Systems, challenges and future directions. *Future Generation Computer Systems*, 160:748–760, 2024.

[Z+24b] Lei Zhou et al. Cloud–edge–end collaborative federated learning: Enhancing model accuracy and privacy in non-iid environments. *Electronics*, 13(24):5021, 2024.

[ZBD+21] Seyed Zekavat, R Michael Buehrer, Gregory D Durgin, Lisandro Lovisolo, Zhonghai Wang, Shu Ting Goh, and Ahmad Ghasemi. An overview on position location: Past, present, future. *International journal of wireless information networks*, 28:45–76, 2021.

[ZLT01] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. Spea2: Improving the strength pareto evolutionary algorithm. *TIK report*, 103, 2001.

[ZQD+20] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.