Literature Review

The usage of distributed computing systems in Machine Learning has been popular in the recent decades. This is due to the increased complexity of the models that arose from the increased usage and popularity of ML algorithms across different fields. Distributed computing systems can be used to efficiently train a neural network that needs an immense number of training hours. In distributed training, the data is partitioned and each partition is assigned to a different machine for processing. The results are then combined to update the model parameters and improve the overall performance.Another way in which distributed computing can be used in machine learning is through distributed modeling. This refers to the use of multiple models to solve a single problem. Each model is trained on a different subset of the data or using a different algorithm, and the results are combined to produce a final prediction. This can help to improve the accuracy of the predictions and reduce the risk of overfitting.In addition to training and modeling, distributed computing can also be used for distributed inference which will be the focus of our work. This refers to the process of making predictions using a pre-trained machine learning model that has been distributed across multiple machines. This can help to improve the speed and scalability of the inference process, particularly for large models or high-volume workloads.

DEFER (Distributed Edge inFERence) [1] (Parthasarathy et al., 2022) is a framework for distributed edge inference that is designed to increase throughput and decrease per-device compute load. The framework partitions deep neural networks into layers that can be spread across multiple compute nodes, with a single "dispatcher" node to distribute the DNN partitions and inference data to the respective compute nodes. The compute nodes are connected in a series pattern where each node's computed result is relayed to the subsequent node. The final result is then returned to the dispatcher node. This approach to distributed edge inference has several advantages. First, it allows the workload to be distributed across multiple compute nodes, reducing the per-device compute load and increasing throughput. Second, by partitioning the DNN into layers that can be processed in parallel, the overall latency of the inference process can be reduced. Third, the series pattern of the compute nodes allows for efficient data relay and reduces the need for complex data synchronization. Overall, DEFER is a promising approach to distributed edge inference that can help to improve the performance and scalability of deep learning applications in edge computing environments. By leveraging the power of distributed computing and parallel processing, DEFER can help to enable a wide range of real-world applications, from autonomous vehicles and drones to smart homes and IoT devices.

The results of the study show that when running the ResNet50 model, the DEFER framework achieves significant performance benefits when compared to single-device inference. Specifically, using DEFER with 8 compute nodes leads to a 53% increase in inference throughput and a 63% reduction in per-node energy consumption.

The paper presented by Zhao et al. (2018) [2] introduces a distributed architecture that offloads the computation of deep learning models to nearby IoT edge devices, and proposes an adaptive strategy for model partitioning and scheduling that dynamically adjusts the allocation of model partitions to the available edge devices based on the current cluster conditions. This framework aims to dynamically balance the load among edge devices and cloud resources to optimize the inference performance of Deep Neural Networks (DNNs).To ensure that our system meets its intended goals, we will analyze and address communication patterns, architectural considerations, and operational challenges. We'll explore solutions to optimize its efficiency and effectiveness, drawing insights from the relevant research papers [3].

The work presented by Lunga et al. (2020) [4] tackles the large data problem when sent over the physical network. Its key strategy is to reduce the amount of data being sent, and this can be done using compression. During the transfer of model and weights between nodes, we could use an encoder to compress data before sending over the physical network to increase bandwidth, before finally decompressing at the destination node.

The paper titled "Distributed Inference with Deep Learning Models across Heterogeneous Edge Devices" proposes a framework for improving the performance of federated learning over heterogeneous wireless networks. The authors discuss the difficulties associated with wireless network heterogeneity and describe three components of their proposed framework: network-aware device selection, model splitting and aggregation, and a communication-efficient data exchange method. The authors use simulations using real-world datasets to show that their framework outperforms previous federated learning algorithms in terms of accuracy and communication cost. Additionally, the article examines the influence of network heterogeneity on federated learning performance and demonstrates that the proposed architecture is resistant to such heterogeneity. In conclusion, this research makes an important addition to the field of federated learning by stressing the significance of taking network heterogeneity into account while creating efficient and resilient federated learning frameworks.[5]

The paper presented by Hu et al. (2022) proposes a new approach for improving the robustness and privacy of federated learning against adversarial data poisoning attacks. The authors begin by highlighting the issues of federated learning and the system's possible vulnerabilities to data poisoning attacks. They then show their suggested technique, which use a multi-objective optimization algorithm to train models in such a manner that the influence of the poisoned data is minimized while still obtaining good accuracy. The authors use simulations to demonstrate that their technique beats previous alternatives in terms of both accuracy and resilience against data poisoning attacks. Additionally, the method is demonstrated to protect the privacy of the training data while still obtaining excellent accuracy. In conclusion, this study contributes significantly to the area of federated learning by tackling the important challenges of robustness and privacy in the face of adversarial data poisoning attacks.[6]

We can take a look at other methods of parallelising such as this paper [7]. Instead of doing it by hand like DEFER, one could opt to use Apache Spark, which is an open-source unified analytics engine for large-scale data processing. It provides an interface for programming clusters with implicit data parallelism and fault tolerance. Furthermore, it was identified that during the model training phase, effective data partitioning allows for more consistent inference over wide geographic conditions. Such a strategy employed in our model could increase the accuracy of our inference.

The paper "Distributed Deep Learning Inference Acceleration using Seamless Collaboration in Edge Computing" investigates inference acceleration in collaborative edge computing utilizing distributed convolutional neural networks (CNNs). While doing segment-based partitioning, we take the receptive-field into account to ensure inference accuracy. To optimize parallelization between communication and computing processes, hence lowering overall inference time of an inference work, we propose HALP, a novel task cooperation method in which the overlapping zone of sub-tasks on secondary edge servers (ESs) is executed on the host ES.We expand HALP to include multiple task scenarios. According to experimental findings, HALP surpasses the state-of-the-art work MoDNN by accelerating CNN inference in VGG-16 by 1.7–2.0x for a single job and 1.7–1.8x for 4 tasks per batch on GTX 1080TI and JETSON AGX Xavier. The service reliability under time-variant channels is also evaluated, and the results demonstrate that HALP is a successful method for ensuring high service dependability under tight service deadlines.[8]

References

[1] A. Parthasarathy and B. Krishnamachari, "DEFER: Distributed Edge Inference for Deep Neural Networks," 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2022, pp. 749-753, doi: 10.1109/COMSNETS53615.2022.9668515.

[2] Zhuoran Zhao, Andreas Gerstlauer. (2018) DeepThings: Distributed Adaptive Deep Learning Inference on Resource-Constrained IoT Edge Clusters. http://slam.ece.utexas.edu/pubs/codes18.DeepThings.pdf

[3] Ivan Rodriguez-Conde, Celso Campos, Florentino Fdez-Riverola. (2023) Horizontally Distributed Inference of Deep Neural Networks for AI-Enabled IoT. https://www.mdpi.com/1424-8220/23/4/1911

[4] Lunga, D., Gerrand, J., Yang, L., Layton, C., & Stewart, R. (2020). Apache Spark accelerated deep learning inference for large scale satellite image analytics. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 271–283. https://doi.org/10.1109/jstars.2019.2959707

[5] C. Hu and B. Li, "Distributed Inference with Deep Learning Models across Heterogeneous Edge Devices," IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, 2022, pp. 330-339, doi: 10.1109/INFOCOM48880.2022.9796896. https://iqua.ece.toronto.edu/papers/chenghao-infocom22.pdf

[6] S. Henna and A. Davy, "Distributed and Collaborative High-Speed Inference Deep Learning for Mobile Edge with Topological Dependencies" in IEEE Transactions on Cloud Computing, vol. 10, no. 02, pp. 821-834, 2022.
https://doi.ieeecomputersociety.org/10.1109/TCC.2020.2978846

[7] [2] Chinchali, S. P., Cidon, E., Pergament, E., Chu, T., & Katti, S. (2018). Neural networks meet physical networks. Proceedings of the 17th ACM Workshop on Hot Topics in Networks. https://doi.org/10.1145/3286062.3286070

[8] N. Li, A. Iosifidis and Q. Zhang, "Distributed Deep Learning Inference Acceleration using Seamless Collaboration in Edge Computing," *ICC 2022 - IEEE International Conference on Communications*, Seoul, Korea, Republic of, 2022, pp. 3667-3672, doi: 10.1109/ICC45855.2022.9839083.