WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Edge AI: Running Machine Learning Models on Edge Devices

Shikha Tiwari[1], Mahak[2]

[1]Assistant Professor, Amity University Chhattisgarh, shkhtiwari583@gmail.com
[2] Student, Amity University Chhattisgarh, nishadmahek02@gmail.com

ABSTRACT

Combining edge computing and artificial intelligence (AI) edge AI allows for data processing and inference near the data source at the edge. This method minimizes bandwidth consumption improves privacy and lowers latency. The demand for effective real-time data processing has grown as Internet of Things (IoT) devices proliferate. Through the direct execution of machine learning (ML) models on edge devices Edge AI overcomes the high latency bandwidth and security issues that plague traditional cloud-based AI systems. Mobile phones Raspberry Pis and NVIDIA Jetsons are examples of devices that can run low-resource-environment-optimized lightweight models. This study examines the importance difficulties designs and practical uses of ML model deployment on edge devices. We also look at upcoming trends like model compression federated learning and how Edge AI will advance with 5G.

Keywords:- Edge Computing, Artificial Intelligence, Machine Learning, Edge AI, Real-Time Inference, Lightweight Models, Federated Learning, 5G Integration, IoT, Model Optimization, Jetson Nano, TensorFlow Lite, On-Device Processing, Low-Latency Systems, Data Privacy.

1.Introduction.

The number of connected devices has increased exponentially as a result of the Internet of Things (IoT) explosive growth producing enormous volumes of data. Even though they are very effective traditional cloud-based AI systems have drawbacks like high latency a need for continuous internet connectivity and privacy issues [1:2]. Time-sensitive applications cannot use these systems because they frequently cannot process data in real-time. By bringing computation closer to the data source edge AI gets around these restrictions and guarantees quicker decision-making less bandwidth consumption and enhanced privacy [3]. Using specialized hardware like the NVIDIA Jetson smartphones and edge devices like Raspberry Pi Edge AI runs machine learning models at the data source [4] [5]. Applications needing low-latency inference like industrial automation healthcare monitoring and driverless cars depend on this decentralization of computation [6]. This study looks at Edge AIs architectures difficulties and practical uses in addition to the upcoming developments influencing its development [7].

2. Architecture for Edge AI.

Sensors edge devices with processing power and machine learning models are the main elements of a typical Edge AI system [8]. Local processing of the data collected by sensors takes place on edge devices that have processors like GPUs TPUs or specialized accelerators. These gadgets are capable of operating lightweight models that have already been trained and optimized for resource and performance limitations. Frameworks like PyTorch Mobile [10] OpenVINO [11] and TensorFlow Lite [9] are made to make it possible to implement AI models on gadgets with constrained processing power. TensorFlow Lite for instance suits mobile and embedded systems by transforming TensorFlow models into a format tailored for edge devices [9]. In contrast PyTorch Mobile provides flexibility for real-time model modifications by enabling dynamic computation graphs [10]. To further optimize AI models for hardware accelerators and edge devices Intel created the OpenVINO (Open Visual Inference and Neural Network Optimization) toolkit [11]. Low-latency inference requires that Edge AI systems architecture take local data processing into account. Real-time decision-making at the edge and highly optimized models are necessary for certain applications that demand quick reactions like autonomous driving or real-time video surveillance [12].

3. Challenges of Operating ML at the Edge

When utilizing machine learning models on edge devices there are several challenges.

- Limited Processing Power and Memory: Edge devices have limited processing power and memory. Complex deep learning models which typically require a lot of GPU or TPU capability are difficult to implement due to these limitations [2] [13]. As a result it is possible to implement only lightweight models such as MobileNets and SqueezeNet which require further optimization through techniques like quantization pruning and knowledge distillation [14] [15].

- Power Usage: A lot of edge devices are utilized in remote or mobile environments with limited power supply. Therefore energy-efficient algorithms and designs that lower energy consumption without compromising performance are needed. Effective resource management, low-power hardware design and model compression are examples of power optimization techniques [16] [17].

- Latency and Model Size: Machine learning models must be smaller in order to take into account the memory constraints of edge devices. Model size can

- be decreased without compromising performance with the use of techniques like model pruning and quantization. Furthermore low-latency inference must be guaranteed especially for real-time applications like autonomous vehicles and industrial automation [18] [19].

4. Real-Life Applications

Edge AI is transforming several sectors by facilitating autonomous systems and real-time decision-making. Some important applications include:

Healthcare: With wearable gadgets like smartwatches and health trackers, Edge AI enables constant vital sign monitoring for patients. These devices have the ability to perform data processing locally and send real-time warnings during emergencies, enabling prompt medical care [1], [4]. Additionally, healthcare practitioners may benefit from early disease detection with the use of AI-powered diagnostic equipment that are used on edge devices [20].

Agriculture: Precision farming uses edge AI to enable the use of drones and soil sensors to monitor crop health and identify diseases in real time. These systems can automatically detect areas that need attention, enabling prompt action and minimizing crop losses [3], [21]. Data may be analyzed locally at the edge, which decreases the demand for expensive data transfer to the cloud.

Autonomous Vehicles: Edge AI is used in autonomous vehicles for functions such as collision avoidance, lane tracking, and object identification. These cars can process sensor data locally, allowing them to make quick judgements without needing cloud connectivity, which increases safety and responsiveness [7], [22].

5. Case Study: Jetson Nano

The Jetson Nano from NVIDIA is a widely used platform for creating edge AI apps. It supports deep learning frameworks like TensorFlow, PyTorch, and Caffe, and it has a 128-core Maxwell GPU. The Jetson Nano is a popular choice among developers and researchers [6], [23] due to its suitability for real-time image classification, object detection, and robotics applications.

In a case study about autonomous robots, the Jetson Nano was used to implement a convolutional neural network (CNN) for real-time object identification and tracking [6]. The robot was able to interact with its environment independently and move through complicated settings thanks to the platform's great computing capability and low latency.

6. Future Trends

A number of new trends will determine the course of Edge AI in the future:

Federated Learning: With federated learning, devices may work together to train models without exchanging raw data. This method is very helpful for healthcare and other privacy-conscious uses where data confidentiality is critical [24], [25].

Model Compression and Optimization: The emphasis will be on optimizing models for efficiency as more potent edge devices are implemented. Approaches such quantization, knowledge distillation, and pruning will be essential for reducing the size and speed of AI models while retaining their accuracy [15], [16].

5G Integration: With its high bandwidth and low latency communications, the introduction of 5G networks will facilitate even greater acceptance of Edge AI. This will improve the performance of real-time apps like remote surgery, self-driving cars, and industrial automation [26], [27].

7. Conclusion

Edge Real-time decision-making is made possible by AI, a quickly developing paradigm that allows for intelligent data processing right at the data source, thereby lowering latency and enhancing privacy.

Despite challenges such as restricted computing resources, energy restrictions, and the necessity for optimized models, ongoing advancements in hardware (e.g., Jetson Nano), lightweight ML frameworks (e.g., TensorFlow Lite), and model optimization methods (e.g., pruning and quantization) are addressing these limitations.

The integration of federated learning, on-device training, and 5G connection will greatly increase the scalability and efficiency of Edge AI systems in the future. The potential uses for Edge AI will grow in sectors like healthcare, autonomous systems, smart cities, and the industrial Internet of Things as a result of these advancements. Consequently, Edge AI is about to play a key role in the development of future intelligent systems.

References

[1] Y. Chen, Y. Hao, and K. Wang, "Edge cognitive computing based smart healthcare system," Future Generation Computer Systems, vol. 86, pp. 403–411, 2018.

https://doi.org/10.1016/j.future.2018.04.029

[2] A. Ghosh, O. G. Ali, and S. S. Gill, "Edge AI: Challenges and opportunities," IEEE Internet of Things Journal, vol. 8, no. 16, pp. 13033–13045, 2021. https://doi.org/10.1109/JIOT.2021.3065135

[3] T. Zhang, Y. Wang, and Y. Wang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," Proceedings of the IEEE, vol. 107, no. 8, pp. 1738–1762, 2019.https://doi.org/10.1109/JPROC.2019.2918951

[4] S. Abedin, M. A. Rahman, N. Hossain et al., "Fog Computing and Mobile Edge Computing Based IoT Infrastructure for Smart Healthcare," Computer Networks, vol. 191, 2021. https://doi.org/10.1016/j.comnet.2021.108040

[5] TensorFlow Lite Documentationhttps://www.tensorflow.org/lite

[6] NVIDIA Jetson Nano Developer Kit https://developer.nvidia.com/embedded/jetson-nano-developer-kit

[7] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7457–7469, 2020. https://doi.org/10.1109/JIOT.2020.2984887

[8] Y. Xie, J. Wang, and D. Ma, "Edge AI: An Overview of Machine Learning on Edge Devices," IEEE Access, vol. 8, pp. 61898–61916, 2020.https://doi.org/10.1109/ACCESS.2020.2989061

[9] P. L. L. D. Jayakumar, "Efficient Edge AI for Real-time Machine Learning at the Edge," Journal of AI Research, vol. 68, pp. 205–220, 2021. https://doi.org/10.1613/jair.6970

[10] A. Garcia, "A Survey of Optimized Edge AI Architectures for Deep Learning," IEEE Transactions on Computational Intelligence and AI in Games, vol. 13, no. 3, AI Assisted Teaching Support for Answering Frequently Asked Questions and Generating Educational Content", 4th IEEE International Conference on Information and Communication Technology in Business, Industry and Government (ICTBIG-2024) held at Symbiosis University of Applied Sciences,Indore, INDIA during 13th – 14th December, 2024.

11. "Blockchain-Powered Decentralized Platforms for Secure Healthcare Data Exchange", 4th IEEE International Conference on Information and Communication Technology in Business, Industry and Government (ICTBIG-2024).

12."AI and Blockchain Synergy for Advanced Health Data Processing in IoT", 4th IEEE International Conference on Information and Communication Technology in Business, Industry and Government (ICTBIG-2024). . "Integrating Deep Learning to Decode Meningeal Interleukin-17 T Cell Mechanisms in Salt-Sensitive Hypertension-Induced Cognitive Impairment." IEEE Explore, https://ieeexplore.ieee.org/xpl/conhome/10685846/proceeding https://doi.org/10.1109/OTCON60325.2024.10687585

13. "Cross-Lingual Transfer Learning in Rnns For Enhancing Linguistic Diversity in Natural Language Processing" 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET) | 979-8-3503-8880-0/24 ©2024 IEEE | DOI: 10.1109/ACROSET62108.2024.10743896