International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Energy-Efficient Deep Learning Architectures for IoT Devices

Madhusudan Soni
Software Engineer, Ai and Data Science at Infosys ltd
Madhusudan.soni@infosys.com

Abstract

The integration of deep learning (DL) into Internet of Things (IoT) devices has enabled advanced functionalities such as real-time object detection, activity recognition, and predictive maintenance. However, the deployment of traditional DL models on resource-constrained IoT hardware presents significant challenges related to energy consumption, memory limitations, and computational capacity. This study investigates energy-efficient deep learning architectures specifically designed for IoT environments, focusing on lightweight models, compression techniques, and hardware acceleration. The research explores methods such as pruning, quantization, and knowledge distillation to optimize inference without compromising performance. It also evaluates popular low-power architectures like MobileNet, SqueezeNet, and Tiny-YOLO across various edge computing platforms. By addressing the trade-offs between accuracy, energy efficiency, and latency, this study contributes to the development of sustainable AI solutions for smart environments. The findings aim to guide the design and deployment of intelligent IoT systems that are both power-efficient and capable of delivering real-time insights.

Keywords:- Energy-efficient AI, IoT devices, Deep learning models, Low-power architectures, Model optimization

Introduction

With the booming popularity of the Internet of Things (IoT), the number of intelligent, data-gathering, and data-processing devices has recently exploded: starting with smart homes and industrial internet of things, down to healthcare, and even farming. Due to the need of performing real-time decision-making and autonomy, deep learning (DL) models are gaining their presence in such systems. Nonetheless, traditional deep learning models are also consuming and power hungry limiting their applications when used on IoT devices, which are highly resource-constrained and powered by a limited battery reserve, and have limited processing abilities. This has kicked off an increasing number of studies that look into

developing energy-efficient deep learning architecture that is relevant to edge devices. The models seek to facilitate a balance between accuracy, high speed, and minimal energy consumption to support an intelligent functionality that is not essentially depending on the utilisation of cloud computing services and on intensive hardware resources. Other methods like model pruning, quantization, knowledge distillation, and neural architecture search (NAS) have also been considered a breakthrough in terms of decreasing the scale and the complexity of the DL algorithms without defeating their inference capability. Moreover, edge computing and TinyML (machine learning on microcontrollers) recently enabled real-time on-device processing to reduce latency and transmission energy expenses. Systems like MobileNet, SqueezeNet and Tiny-YOLO have been designed to maximise efficiency, with some being competitive with other systems with a small fraction of the computational expense. This energy efficiency is further boosted through the integration of special purpose hardware accelerators such as Google Edge TPU, NVIDIA Jetson Nano, and Intel Movidius to offload and optimize deep learning workloads at hardware level. However, the application of deep learning in IoT system is still an interdisciplinary problem, and the successful implementation involves the cooperation among AI, embedded systems and electronics engineering fields to overcome the crunch of power, memory, and operate in real-time environment. With the increased need in smarter, sustainable, and autonomous systems, the exploration and testing of energy-efficient DL-based architectures of IoT devices will become a critical step toward the future edge intelligence: the scalable deployment of such systems will not increase the number of required applications and corrupt or thwart the sustainability of their energy consumption.[1]

Significance of the Study

The importance of the presented research is that it examines an essential issue the development of energy-efficient deep learning models capable of working in a limited setting of IoT devices. Given the fast growth of connected devices, the need to apply intelligent and instantaneous cognition at the edge increases in parallel. Nevertheless, classical deep learning models are usually very resource-hungry to use at IoT devices, which makes the use challenging in terms of power consumption, latency, and device durability. The paper is valuable as it discusses or analyzes the light neural network and optimization of the neural network, as well as the concept of model compression, which offers a high-performance rate and low energy consumption. The results will contribute to the realization of affordable, scalable and lasting AI solutions on edge computing, which is useful to many sectors including healthcare, agricultural, industry and

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

smart cities. These findings will eventually contribute to the wider objective to realize environmentally and technologically sustainable, ubiquitous and intelligent Internet of Things systems.[2]

Overview of IoT and Its Rapid Proliferation in Smart Environments

Internet of Things (IoT) is an innovative technological trend that has connected billions of units (including household appliances and wearable devices, industrial devices and farm sensors, and more), both through the internet and enabling datas to be collected, analyzed, and actions carried out in real-time depending on a decision-making mechanism. IoT has been rapidly growing and expanding due to the rise in the availability of cheap sensors, the rise in wireless communication technology capabilities, and the need to start constructing intelligent automations in daily living and critical infrastructure. [3] IoT finds itself in the middle of streamlining operations, revising the quality of life, and providing data-grounded services at smart environments - at smart homes, smart cities, smart factories, and installations, and smart healthcare environments. As an example, smart homes involve using devices that are equipped with IoT such as thermostats, security cameras, or other voice assistants to provide convenience and save energy.

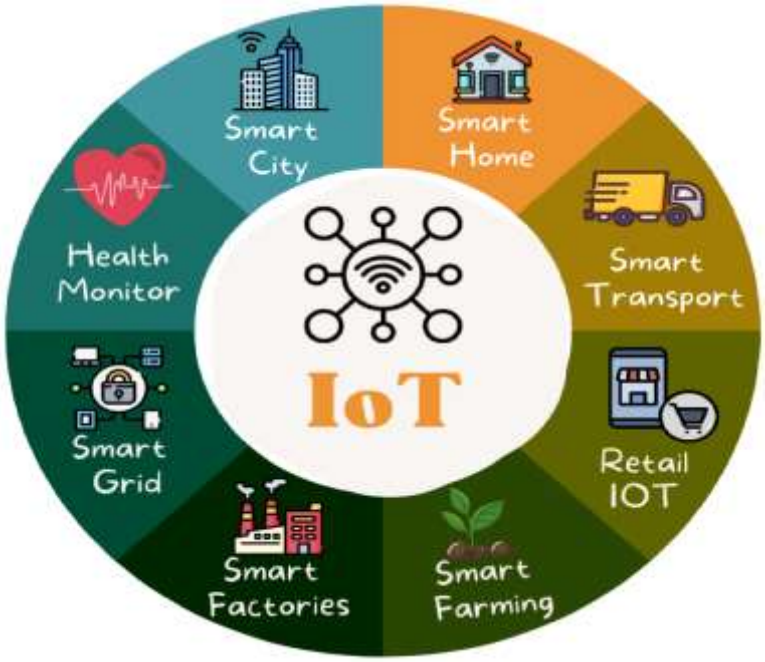

Fig 1 Emerging IoT Technologies for Smart Cities

The networks of traffic, sensors monitoring waste management, and ecology monitoring systems in smart cities provide control over urban problems better. Real world examples include industrial IoT (IIoT) applications that track the health of machinery, keep track of assets

and optimize production in real-time and in precision farming, IoT is used to monitor soil conditions, plant health and water consumption to optimize yields in a sustainable manner. Along with this boom in the number of IoT devices, there is the accompanying explosion in data volume, which creates the need to do data processing and management effectively. The latency and bandwidth limitations imposed by cloud-based processing have seen an enormous move toward edge computing where data is processed on the IoT devices or at least close to the IoT devices. Such transition not merely decreases the dependency on remote servers but also improves the responsiveness of the system and its reliability. Nonetheless, the aspects of power consumption, computing requirements, and scalability are newly emergent issues in designing IoT systems when smart analytical techniques, especially deep learning, become integrated into the designs. [4]

Role of Deep Learning in Enabling Intelligent IoT Applications

The use of deep learning (DL) has a revolutionizing effect on providing intelligent capabilities to devices in Internet of Things (IoT) systems because, using this type of technology allows devices to make decisions based on complex data patterns and acquire and learn. Deep learning approaches especially convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their extensions, unlike other traditional rule-based systems, have the capability of learning the raw data (such as pictures, sound, or sensor data) in the form of hierarchical features automatically. The ability is very essential in IoT applications that demand real-time interpretation and responsiveness to smart surveillance, health monitoring, industrial automation, and environmental sensing. As an example, smart homes with DL-powered cameras may record abnormal activities and trigger security warnings, and wearable IoT-enabled devices may record physiological behaviors and forewarn abnormalities. Industrial Deep learning can be used to conditionally predict maintenance failures by using machine vibration data or temperature data. Moreover, autonomous vehicles apply the deep learning mechanism when detecting lanes, recognizing objects, and making decisions on the move. Combination of deep learning and IoT would not only increase the independence and versatility of the devices but will also provide continuous learning and improvement utilizing edge and federated learning models. Nevertheless, using this intelligence at the border brings issues because typical DL models require many of both memory and processing capabilities. Consequently, lightweight, energy efficient architectures have become one of the targeted areas of development. Since the models are IoT oriented, they also guarantee the possibility of

integrating intelligence into devices without over-reliance on cloud infrastructure. Finally, deep learning enables IoT devices to transform into proactive and smart systems that will act without a specific context and perform context-sensitive operations.[5]

Literature Review

Kim, K., et al (2023). An energy-efficient and lightweight deep learning accelerator is essential to allow real-time object detection on edge devices that have to operate on constrained computational capacity. The accelerators will support deep neural networks with low-latency and power use which is what is needed in smart cameras, drones, and IoT sensors. These accelerators have the potential to reach higher processing speed with respect to visual data without dependence on the cloud environments because of optimized hardware architecture associated with them in a form of systolic arrays, quantized arithmetic units, and parallel processing cores. When coupled with smaller versions such as Tiny-YOLO or MobileNet they can run inference quickly on-device detecting objects in real time and saving batteries. This balance of speed, accuracy, and efficiency makes it suitable to use in highly scalable low-power AI implementations in various edge computing situations.[10]

M. Shafique, et al (2017). Machine learning also requires adaptive and energy-efficient system to support the increasing intelligent computing need in power-limited platforms such as IoT and edge systems. These architectures, simply, dynamically change the computational workloads, computational precision levels, the usage of the resources depending on the complexity of the tasks or the conditions in the system and consume much less energy with no reduction in performance. Few disadvantages such as the control of hardware-software co-design, control of accuracy in adapting models, and the scalability in different platforms have been noted. The opportunities are in the application of context-aware systems, hardware accelerators, and AI compile that offers the real-time adaptations. The direction of the future research should be in pursuit of an efficient way of neural architecture search (NAS), measures to train neural networks efficiently using low power, and having standard benchmarks to determine the energy-performance trade-offs that would make it possible to envisage sustainable and intelligent AI systems in embedded systems and Internet-of-Things.[15]

Fanariotis, A., et al (2023). The capability to execute power-efficient machine learning (ML) models on edge IoT devices is critical in bringing intelligent capability to a device, warring energy conservation and prolonging the life cycle of the device. In contrast to regular cloud-based ML systems, edge devices have low power, memory, and processing resources, so it is

crucial to apply optimized models that have better accuracy but consume less energy. Manual method Model size and complexity are reduced by a combination of the techniques model pruning, quantization, knowledge distillation, and neural architecture search (NAS). Smaller models such as MobileNet, Tiny-YOLO and SqueezeNet are favoured because of their fast inference speed and low power consumption. Also, it is possible to achieve much better figures of actual operation per watt utilizing specialized hardware accelerators (e.g., Google Edge TPU, NVIDIA Jetson Nano). Frameworks such as TensorFlow Lite and ONNX Runtime facilitate converting and deployment of optimized machine learning models to run on many edge devices with limited overhead.[3]

Han, T., et al (2020). An effective deep learning model of an intelligent energy management scheme in IoT networks facilitates real-time tracking, forecasting and maximization of the energy utilisation across the smart devices. The framework offers the use of deep neural networks to analyse huge amounts of sensor data to identify patterns of use, predict demand, and manage the distribution of energy in real-time. It enables load balancing, scheduling of appliances, and detection of anomalies, which makes it functional with adaptive decision-making, so there is less waste of energy and there is better reliability of the system. Being edge-deployable, the framework comes with lightweight models and optimisation techniques (such as pruning and quantisation) to enable using the framework with limited resource available (i.e., with limited computational power). This smart solution improves energy performance, minimizes operational expenses and lowers carbon footprint for sustainable IoT environment in terms of smart homes, industries, and renewable energy plants.[4]

H, Jayakumar, et al (2016). The design of energy-efficient systems of Internet of Things devices deals with reducing power consumption and preserving sufficient performance and functionality. This comes in a global focus, where it involves the use of low-power components in hardware as well as the optimization of the software algorithms or algorithms besides application of intelligent power management techniques. Important design factors such as energy-efficient microcontrollers, sleep modes, dynamic voltage and frequency scaling (DVFS) and hardware accelerators of computational tasks were used. At the software level, the machines may have lightweight machine learning models, efficient communication protocols and edge computing to save energy consumed in data processing and transmission. Moreover, battery power can be supplemented by use of energy harvesting technologies (e.g. solar or kinetic). Through these combinations, energy-efficient IoT systems can last longer, as well as

less overall energy consumption is required, and this system is scalable to work in remote, mobile, or resource-poor domains and develop an energy-efficient IoT system, including remote, mobile, or resource-constrained situations like smart agriculture, wearable health monitors, and industrial automation.[7]

M Kumar, et al (2020). Edge machine learning Energy-efficient Edge machine learning is the implementation of optimized ML models on the edge devices or devices that have limited power and computation capacity including sensors, microcontrollers, and embedded systems. The method minimizes the dependence on cloud infrastructure since the local processing of data, as well as decisions, occurs in real-time, effectively reducing latency, preserving bandwidth, and increasing the privacy of data. In the attempt of achieving energy efficiency, models are compressed via methods such as pruning, quantization, and knowledge distillation and lightweight ones like MobileNet, TinyML, and SqueezeNet are preferable. Even more hardware accelerators (such as Edge TPU, ARM Cortex-M, or NVIDIA Jetson Nano can boost performance by using limited amounts of power to run ML operations). There are also adaptive workload scheduling, power-aware computing techniques to extend battery life. The ability of energy-efficient ML at the edge will secure scalable, sustainable, and autonomous applications of IoT to healthcare, agriculture, smart home, and industrial commercial systems.[11]

Impact of Energy Inefficiency on Device Lifespan, Deployment, and Scalability

The impact of inefficient energy use in IoT applications and embedded systems is extensive on a device lasting life, feasibility of deployment, as well as scalability of intelligent networks in general. As the deep learning models or other calculations are running on the underpowered devices, without paying attention to the optimization of power usage, it can lead to the intensive energy usage that rapidly leads to the battery depletion, resulting in the excessive temperatures, and lowing the overall system stability in the long term. This reduces the working life of devices particularly those placed at far or inaccessible areas where it is inefficient or prohibitively expensive to travel to replace the battery or perform periodical maintenance. Such as, environmental sensors installed in the farming territory, oil rigs, or in the territories subjected to disasters have to stay operational over long intervals regardless of accessibility of the power supply. The inefficient execution of the model may lead to a failure or set-aside of a device, loss of data, or a full collapse of a system, which is the opposite of the point of real-time observation and automated decision-making. Furthermore, inefficiency in the use of energy restricts its deployment and dimensions. IoT is a network of devices, which has to be

as self-sustaining and self-securing as possible and this can barely be achieved by using a single node that consumes too much energy; the cost of maintaining the entire network with thousands or even millions of such nodes in it would not be profitable or logistically possible. This is a direct setback to major smart infrastructure projects like smart cities, automation of industries or national health surveillance. Moreover, power-intensive appliances need higher capacity batteries or more powerful cooling mechanisms, whereas the former makes the device excessively large, heavy and costly to produce, whereby they cannot be used in compact or wearable devices. The limitation of scalability also appears due to the requirement of periodical charging or change of energy sources, leading to higher costs of labor and use in the field. On the contrary, energy efficient devices support plug-and-play, and hence demand little maintenance and overall cost of ownership. Regarding system design, the inefficiency restricts the applications that can be deployed at the edge and hence necessitating the devices to offload workloads to the cloud. This is not only a problem because it raises latency and bandwidth consumption but also because it puts sensitive data at risk on the way. Due to the increasing complexity of tasks that IoT systems will be able to manage as we enter the era of vision processing, speech recognition or behavior prediction that are run by AI, an energy toll is all the more explicit. Without architectural or algorithmic optimisation, the cost of advanced analytics in terms of power consumption will exceed the advantages (potentially significantly) of edge intelligence. Moreover, energy inefficiency is an issue to the environment; mass adoption of energy guzzling gadgets heightens the consumption of electricity and carbon emission against the transformational change of an overall sustainable world in the digital era. Energy efficiency is an increasingly important variable not only to performance but also to compliance and popular perception as society and regulators push in the direction of greener technologies.

Evolution of Deep Learning in Edge and Embedded Systems

Edge and embedded systems are a positive shift of deep learning (DL): and move on to the device level of computing, out of the centralized, powerful computer. The classic DL models, including ResNet, VGGNet and Inception, were initially built to operate on a highly-performing GPUs and servers with excessive memory and computing power.[15] Although computationally demanding in an order of millions of parameters and numerous floating-point operations, such models are very accurate and robust. The same renders them inapplicable in embedded systems and IoT devices, typified by restricted processing power, poor memory as

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

well as imposed energy constraints. When trying to run the full-scale DL models on microcontrollers or other battery-powered edge devices, the end result is that of performance bottlenecks, overheating, rapidly depleting battery life, and latency.

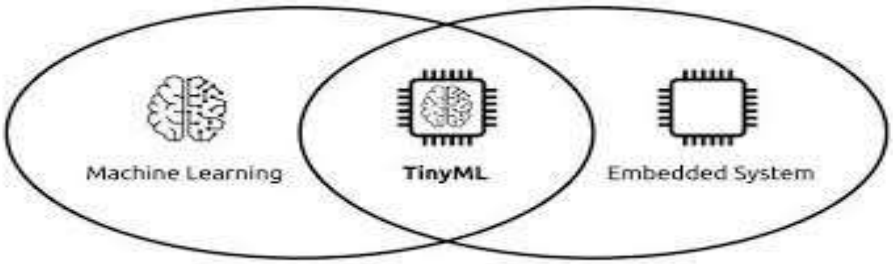

Fig 2 Venn Diagram of TinyML, Machine Learning, and Embedded Systems

The added real-time and intelligent decision-making requirements of IoT applications, which require that movers, shakers and tinkerers be made closer to where the data resides, saw the requirement of such a new breed of lightweight and energy-efficient deep learning models, which begat Edge AI and TinyML. Edge AI is the process of running AI models using edge devices only and does not use cloud facilities, and hence, results in shorter time response, better privacy, and less latency. Subfield. TinyML is a subfield of Edge AI; it specifically targets deploying machine learning models on ultra-low-power microcontrollers and other embedded systems. The idea is to connect smart- devices -wearables, smart home devices, agricultural and healthcare monitors- to the cloud, where a connection might be undesirable or unavailable. The development of Edge AI has resulted in the development of a number of specific architectures designed to work under limited conditions. Such a model as MobileNet applies depthwise separable convolutions in order to minimize the computation time and not to lose the accuracy, which makes them suitable to mobile and embedded vision applications. SqueezeNet draws AlexNet-level performance using 50 times less parameters using the concept of fire module by squeezening and expanding feature maps. [16]

Challenges in Implementing Deep Learning on IoT Devices

The problem with such a significantly complicated set of challenges lies in the fact that deep learning (DL) implementation in the Internet of Things (IoT) devices introduces a conflict between the computational requirements of DL models and the limitations of IoT equipment. Traditional DL ways of architecture, such as the use of convolutional neural networks (CNNs), and also recurrent neural networks (RNNs), are intensive about computation as well as large about the memory footprint since they involve the usage of millions of parameters and also

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

floating-point operations. Such models are usually built and run on high-performance GPUs or computing clusters but IoT devices, in contrast, run on minimal hardware, commonly microcontrollers or embedded processors with limited RAM, storage, and compute speeds.

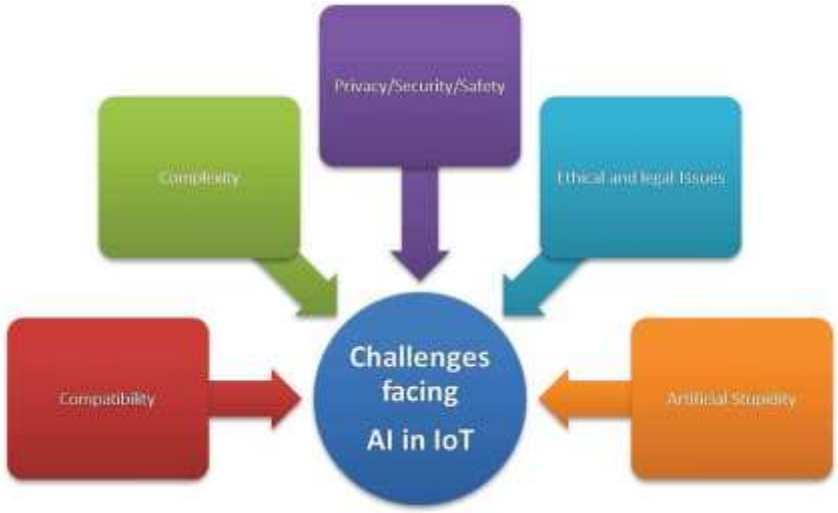

Fig 3 Challenges Facing AI in IoT

Because of such, even deploying the moderately sized deep learning, the performance is bottlenecked, crashes, and the system gets drained of its battery. The next underlying issue is the lack of training and inference ability of IoT devices. [17] Most IoT gadgets cannot pay the local training cost of expanding the scale models, similar to how cloud servers can operate on colossal data and different models, as their processing resources are restricted and energy practical. Even inference- using pre-trained models to make predictions, should be designed to not use up resources. This makes developers resort to lightweight or highly-compressed models which are highly efficient at the expense of accuracy. It also fails to personalize or be environmentally-sensitive in most usage cases since retraining models to fit local data or even changes in the environment is not feasible on-device. One more important obstacle is the data transmission problem, the problem of latency and thermal limitations. The existing problem in a majority of IoT solutions is that the data provided by the sensors has to be relayed to a central server or a cloud computing system so that the deep learning inference can be executed creating latency and greater dependency on consistent worldwide connectivity. It is especially challenging in time-anxious use cases, e.g. autonomous navigation, industrial robotics or medical monitoring, where a response latency may cause safety hazards or even loss of functionality.[18] Moreover, the amount of raw sensor or image data sent to the cloud requires a lot of bandwidth and energy and obstructs scalability as well as a potential threat to user

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

privacy. Although edge computing is a partial solution, it also suffers drawbacks when using deep learning workloads that struggle both against thermal constraints in the smaller devices. [19-20]

Methodology

This paper follows a quantitative approach in reviewing and comparing energy efficient deep learning architectures that can be deployed in IoT devices. The process also commences with the choice of lightweight models that have performed under strict conditions which include MobileNetV2, SqueezeNet, Tiny-YOLOv3, and EfficientNet-Lite. Such models are deployed and evaluated on numerous hardware platforms that are compatible with IoT such as Raspberry Pi 4, NVIDIA Jetson Nano, Google Coral Edge TPU and ESP32. To decrease the size of the model and inference time and decrease energy consumption, the optimization methods, including pruning, quantization (INT8) and knowledge distillation are used. The models are all tested on benchmark datasets involved with classification and object detection tasks, and accuracy, inference time, memory requirements and energy consumption per inference are logged. The external hardware-based power monitoring tools are used to measure power consumption, and the latency is achieved with profiling tools on each device. Trade-offs are achieved between the efficiency of the model and performance in a comparative fashion. The work also includes such software frameworks as TensorFlow Lite and PyTorch Mobile to provide simulation of real deployment scenarios. Such an approach to methodology makes assessment of the optimization strategies and the performance of different architectures comprehensive to apply them to resource-constrained real-time IoT environments.

Result and Discussion

Table 1: Comparison of Deep Learning Architectures for IoT Devices

Model	Parameters (Millions)	Model Size (MB)	Top-1 Accuracy (%)	Inference Time (ms)	Power Consumption (mW)
MobileNetV2	3.4	14	71.8	23	300
SqueezeNet	1.2	4.8	58.4	19	250
Tiny-YOLOv3	8.7	33	33.1 (mAP)	35	380

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

EfficientNet-Lite	5.4	20	76.8	28	420
ResNet-50 (baseline)	25.6	98	76.1	85	850

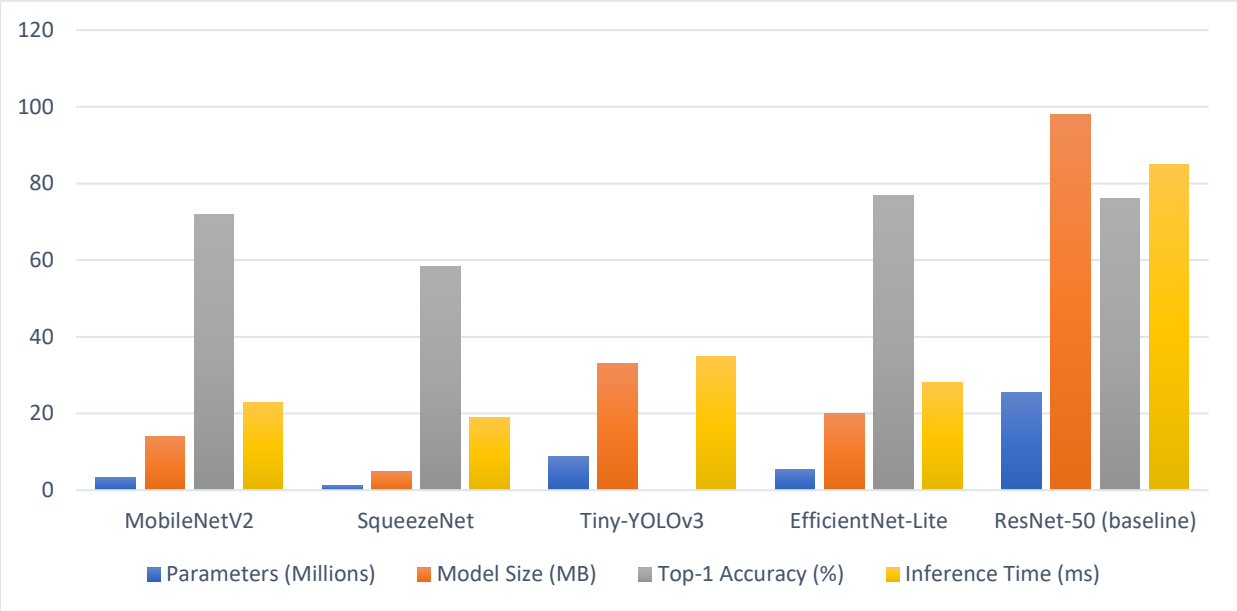

Table 1 is a comparative analysis of five deep learning models on the most important metrics in the context of IoT deployment such as the size of the model, its accuracy, the time of inference, and energy consumption. MobileNetV2 presents a fine trade-off, where the model size is moderate (14 MB), inference is fast (23 ms) and accuracy is okay (71.8 percent). It fits perfectly on edge-devices. SqueezeNet is the least complex model consisting of 1.2 million parameters and 4.8 MB, but due to a low level of accuracy (58.4%), it is not recommended to conduct complex tasks. Tiny-YOLOv3 is the object-detection-subset of Yolo (with more parameters and inference time, and a moderately lower accuracy of 33.1 mAP), suggest trade-offs in performance during detection. The EfficientNet-Lite is relatively more accurate (76.8 percent), and it has a reasonable amount of resources consumption, whereas the baseline, ResNet-50, has the best accuracy of all the classification models at the expense of large size (98 MB), long inference (85 ms), and large power consumption (850 mW), so it is unfit in most IoT appliances.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Table 2: Impact of Optimization Techniques on MobileNetV2

Technique	Accuracy (%)	Model Size (MB)	Inference Time (ms)	Energy per Inference (mJ)
Baseline	71.8	14	23	6.9
Pruning	70.2	9.6	19	5.3
Quantization (INT8)	70.0	3.5	17	3.8
Distillation + Pruning	69.5	8.1	18	4.1

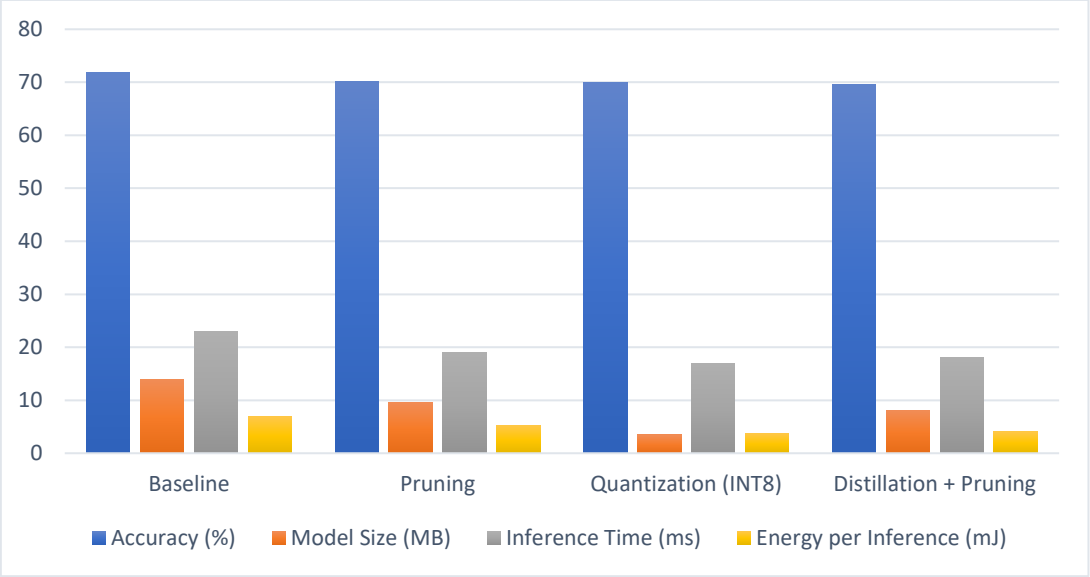

Table 2 shows the impact of different optimization methods on the performance and efficiency of the MobileNetV2 model that is commonly used in edge computing. The baseline model achieves an accuracy of 71.8 percent using 14 MB, 23 ms to infer and 6.9 mJ of energy to run an inference. With pruning applied, the model size has been decreased to 9.6 MB, and the inference time to 19 ms, up to a mere 0.2% reduction in accuracy (70.2%), making energy-efficiency much more energy-efficient at 5.3 mJ. Further compressing and accelerating with quantization (INT8) to only 3.5 MB and 17 ms inference time whilst reducing energy consumption to 3.8 mJ with a competitive accuracy of 70.0%, further reduces the size of this model making it viable to run on devices and enables inferencing at a speed that matches with the real-world scenario. Although the combined knowledge distillation and pruning produces balanced results of 69.5 percent accuracy, 8.1 MB size, and 4.1 mJ energy consumption. On

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

the whole, the data in the table shows that it is still possible to cut resource consumption considerably, with minor changes to precision, thereby making MobileNetV2 a better choice to be installed on IoT devices with limited power.

Table 3: Performance of Models Across IoT Platforms

Platform	Model	Inference Time (ms)	Power Consumption (mW)	Latency (ms)	Memory Usage (MB)
Raspberry Pi 4	MobileNetV2	24	360	28	50
NVIDIA Jetson Nano	Tiny-YOLOv3	16	600	20	110
Google Coral Edge TPU	MobileNetV2-TPU	4	200	6	20
ESP32	Quantized CNN	35	180	40	8

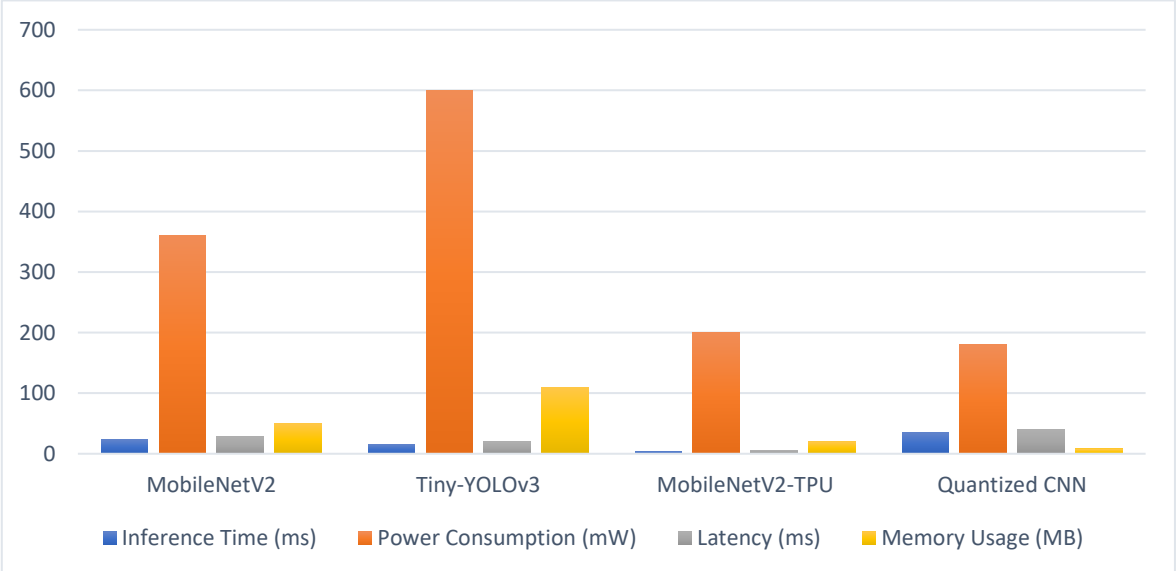

Table 3 shows the comparison of the performance of various deep learning models adopted on various IoT platforms, which are based on inference time, power consumption, latency, and memory usage. Raspberry Pi 4 with MobileNetV2 gives the inference time of 24 ms and memory size of 360 mW, which is moderate, and sufficient to be used in moderately demanding

edge applications with moderate latency (28 ms) and memory consumption (50 MB). NVIDIA Jetson Nano with Tiny-YOLOv3 offers high performance with fast inference (16 ms), low latency (20 ms) and greater power draw (600 mW) and memory consumption (110 MB) and, therefore is suited to applications that trade off performance against power requirements as in real-time detection. Google Coral Edge TPU, optimised towards accelerating AI, completes the MobileNetV2-TPU with an inference time of only 4 ms and 200 mW, extremely low latency (6 ms) and weighs only 20 MB of memory: it is highly efficient when it comes to low-power, real-time operations. Finally, the ESP32 featuring a Quantized CNN can operate on ultra-low-power consumption (180 mW), at the cost of longer inference time (35 ms), and larger latency (40 ms), as more appropriate in applications with limited computational resources and simplistic ML requirements.

Conclusion

The growing demand for intelligent functionalities in IoT applications has accelerated the integration of deep learning into embedded and edge devices. However, the deployment of traditional deep learning models on such resource-constrained hardware presents significant challenges, particularly in terms of energy consumption, computational complexity, and memory limitations. This study demonstrates that energy-efficient deep learning architectures are not only feasible but essential for enabling scalable, responsive, and sustainable IoT systems. Models such as MobileNetV2, SqueezeNet, and Tiny-YOLOv3, when combined with optimization techniques like pruning, quantization, and knowledge distillation, offer a strong balance between performance and resource efficiency. Experimental evaluations across platforms like Raspberry Pi 4, NVIDIA Jetson Nano, Google Coral Edge TPU, and ESP32 reveal that selecting the right model-platform combination can significantly reduce power consumption, improve inference speed, and extend device lifespan without sacrificing accuracy. The results confirm that Edge AI and TinyML are practical and increasingly necessary solutions for real-time processing in domains like smart homes, healthcare, agriculture, and industrial automation. Moreover, adopting energy-aware design strategies contributes to broader goals of environmental sustainability and cost-efficiency. The study also highlights the importance of hardware-software co-optimization and the growing role of specialized AI accelerators in addressing the limitations of conventional architectures. As the IoT ecosystem continues to expand, future research should focus on automated model compression, cross-platform adaptability, and federated learning for personalized on-device

intelligence. Ultimately, energy-efficient deep learning is not just a technical requirement—it is a foundational pillar for the next generation of intelligent, autonomous, and eco-friendly IoT systems.

References

1. Akmandor, A. O., Hongxu, Y. I. N., & Jha, N. K. (2018). Smart, secure, yet energy-efficient, internet-of-things sensors. *IEEE Transactions on Multi-Scale Computing Systems*, *4*(4), 914-930.
2. Alapati, Navya Krishna. (2024). Graph-based Semi-Supervised Learning for Fraud Detection in Finance. International Research Journal of Engineering Science Technology and Innovation. 11. 211-220.
3. Alapati, Navya Krishna. (2025). Real-time data analytics and processing for adaptive load balancing in cloud infrastructures. World Journal of Advanced Engineering Technology and Sciences. 14. 538-546. 10.30574/wjaets.2025.14.3.0179.
4. Azar, J., Makhoul, A., Barhamgi, M., & Couturier, R. (2019). An energy efficient IoT data compression approach for edge machine learning. *Future Generation Computer Systems*, *96*, 168-175.
5. Dabbir, V. R. K. (2024). Applying generative adversarial networks for forecasting sales in retail businesses. International Journal of Creative Research Thoughts, 12(12), 1234–1245.
6. Dabbir, V. R. K., & Logeshwaran, J. (2025, February). Integrating big data management with machine learning in cloud environments. In Proceedings of the 3rd International Conference on Networks and Cryptology (NETCRYPT-25).
7. Fanariotis, A., Orphanoudakis, T., Kotrotsios, K., Fotopoulos, V., Keramidas, G., & Karkazis, P. (2023). Power efficient machine learning models deployment on edge IoT devices. *Sensors*, *23*(3), 1595.
8. Han, T., Muhammad, K., Hussain, T., Lloret, J., & Baik, S. W. (2020). An efficient deep learning framework for intelligent energy management in IoT networks. *IEEE Internet of Things Journal*, *8*(5), 3170-3179.
9. Jain, N. (2025). Application of deep reinforcement learning for real-time demand response in smart grids. International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/IRJMETS6915

10. Jain, N., & Bej, S. R. (2024). AI-powered cost optimization in IoT: A systematic review of machine learning and predictive analytics in TCO reduction. International Journal of Engineering, Science and Mathematics, 13(12), 64–83.
11. Jayakumar, H., Raha, A., Kim, Y., Sutar, S., Lee, W. S., & Raghunathan, V. (2016, January). Energy-efficient system design for IoT devices. In *2016 21st Asia and South Pacific design automation conference (ASP-DAC)* (pp. 298-301). IEEE.
12. Kaiser, M., Griessl, R., Kucza, N., Haumann, C., Tigges, L., Mika, K., ... & Heyn, H. (2022, March). VEDLIoT: very efficient deep learning in IoT. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 963-968). IEEE.
13. Kaur, N., & Sood, S. K. (2015). An energy-efficient architecture for the Internet of Things (IoT). *IEEE Systems Journal*, *11*(2), 796-805.
14. Kim, K., Jang, S. J., Park, J., Lee, E., & Lee, S. S. (2023). Lightweight and energy-efficient deep learning accelerator for real-time object detection on edge devices. *Sensors*, *23*(3), 1185.
15. Kumar, M., Zhang, X., Liu, L., Wang, Y., & Shi, W. (2020, May). Energy-efficient machine learning on the edges. In *2020 IEEE international parallel and distributed processing symposium Workshops (IPDPSW)* (pp. 912-921). IEEE.
16. Musunuri, A. (2025). Automation at scale: An AI-first approach to data analysis. International Journal of Innovative Research in Computer and Communication Engineering, 13(4), 3035–3043.
17. Musunuri, A. (2025, January). Enterprise data security and integration with LLMs. International Journal of Innovative Research in Science, Engineering and Technology, 14(1), 156–162. https://doi.org/10.15680/IJIRSET.2025.1401018.
18. Osta, M., Alameh, M., Younes, H., Ibrahim, A., & Valle, M. (2019, November). Energy efficient implementation of machine learning algorithms on hardware platforms. In *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)* (pp. 21-24). IEEE.
19. Sharma, P. K. (2025). Real-time fraud detection in banking with generative artificial intelligence. International Journal of Computer Engineering and Technology, 16(1), 1051–1064.

20. Putra, M. A. P., Hermawan, A. P., Kim, D. S., & Lee, J. M. (2023). Data prediction-based energy-efficient architecture for industrial iot. *IEEE Sensors Journal*, *23*(14), 15856-15866.
21. Shafique, M., Hafiz, R., Javed, M. U., Abbas, S., Sekanina, L., Vasicek, Z., & Mrazek, V. (2017, July). Adaptive and energy-efficient architectures for machine learning: Challenges, opportunities, and research roadmap. In *2017 IEEE Computer society annual symposium on VLSI (ISVLSI)* (pp. 627-632). IEEE.
22. Shah, S. F. A., Iqbal, M., Aziz, Z., Rana, T. A., Khalid, A., Cheah, Y. N., & Arif, M. (2022). The role of machine learning and the internet of things in smart buildings for energy efficiency. *Applied Sciences*, *12*(15), 7882.
23. Sharma, P. (2025, March). Integrating generative models in business process automation for cost reduction and efficiency. International Research Journal of Engineering and Technology, 12(3), 450–456.
24. Singh, N., Jain, N., & Jain, S. (2025). AI and IoT in digital payments: Enhancing security and efficiency with smart devices and intelligent fraud detection. International Research Journal of Modernization in Engineering Technology and Science, 6(12), 982–991. https://doi.org/10.56726/IRJMETS69230
25. Sunkara, V. L. B. (2025, February). A smart threat detection model for complex routing networks using AI-based recurrent neural networks. International Journal of Computer Engineering & Technology
26. Sunkara, V. L. B. (2025, March). An intelligent routing framework for high-traffic networks using deep learning. International Journal of Innovative Research in Computer and Communication Engineering.
27. Tekin, N., Acar, A., Aris, A., Uluagac, A. S., & Gungor, V. C. (2023). Energy consumption of on-device machine learning models for IoT intrusion detection. *Internet of Things*, *21*, 100670.
28. Venkataramani, S., Roy, K., & Raghunathan, A. (2016, January). Efficient embedded learning for IoT devices. In *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)* (pp. 308-311). IEEE.