Artificial intelligence in the Internet of Things: Integrating and optimizing AI algorithms for real-time data processing and decision-making

Pan Yuan

University of Sydney, Sydney, Australia

2285985550@qq.com

Abstract. The introduction of Artificial Intelligence (AI) will provide great opportunities and pose significant challenges in Internet of Things (IoT) systems. IoT devices can create huge amount of data in real-time. Meanwhile, it is crucial to handle immediate processing and decision-making in IoT applications. However, IoT devices are usually constrained by limited computation power, memory, and energy compared to traditional devices, making the deployment of efficient AI algorithms a challenging task. In this paper, we present different algorithmic strategies to overcome the above problems, including model compression, quantization, and hardware accelerators. On the other hand, we also focus on the role of decentralization and edge computing architectures to increase the scalability and performance of AI-IoT systems. Finally, this paper also reviews energy-efficient AI algorithms and latency reduction methods to make the decision process real-time for various IoT applications.

Keywords: Artificial Intelligence, Internet of Things, AI Optimization, Real-Time Decision-Making, Edge Computing.

1. Introduction

The massive growth of Internet-of-Things (IoT) devices has led to interconnectivity of billions of devices and generated huge amounts of data, which has to be processed and analyzed in real-time. The rapid growth of IoT has been a significant development but there is a huge challenge of processing the tremendous amount of data generated by IoT systems. The volume and velocity of data generated by IoT systems and the ability of AI to process the data make them highly dependent on each other. The complexity of IoT systems may become more sophisticated with the interconnection of multiple systems, where the amount of data will also increase. The current paradigm of learning for AI is too slow for real-time processing that is necessary for IoT systems. Many systems e.g., weather forecasting, predictive policing, ambulance movement, and emergency calls cannot afford delay in processing and need instant decisions. The IoT devices are often resource-constrained as they have either limited computational power, memory, and battery, or a combination of these challenges. Many cloud and information-intensive IoT applications are faced with the challenges of such constraints. Hence, there is a need to resolve these issues [1]. This paper investigates the potential of optimising AI algorithms and seamlessly integrating them into IoT systems in order to process data in real-time and make instantaneous decisions.

Development of AI-driven IoT device poses several challenges, mainly due to resource-constrained nature of AI-driven IoT devices. These devices may have limited computational capabilities, memory and energy, thus embedding AI in them requires overcoming significant hardware and software barriers. These include optimising AI algorithms for execution on power-constrained processors, designing data management systems that can handle large amounts of data generated by IoT, and minimising latency to enable real-time decision-making. This problem becomes more severe when the AI-driven IoT system is scaled across large-scale networks of IoT devices. The main challenge in development of robust IoT AI systems can be traced back to the need to optimise algorithms such that it performs well with limited computational and hardware resources, while not compromising on the accuracy and efficiency.

This paper investigates and proposes solutions to the challenges of integration and optimisation of AI algorithms in the IoT. The first objective is to identify and analyse the hardware constraints affecting AI implementation in IoT devices and propose optimisation techniques and frameworks to cope with the issues arisen. The research also addresses the following aspects: 1) Software architectures that enable AI humanisation and power the functioning of AI-driven IoT, focusing on the use of distributed and edge computing approaches. 2) Data management in IoT and the challenges that it raises regarding AI integration. 3) Energy-efficient AI-driven algorithms and techniques for reducing latency. 4) Approaches to scaling AI-driven IoT and ensuring robustness and efficiency in processing real-time data and making decisions.

2. Integration of AI Algorithms in IoT Devices

2.1. Hardware Constraints and AI Implementation

Using AI algorithms on IoT devices has its own set of challenges due to the hardware limitations of these devices. In general, IoT devices are designed to be small, low-cost and low-power devices – implying their computing power and memory capacity are limited accordingly. This is a great challenge for AI, since many AI algorithms require considerable amount of computational power and memory capacity to function. To address these challenges, various optimisation techniques have already been developed to make AI algorithms more adaptable to resource-limited environments. One way is to use model compression techniques, which reduce the size of the AI model by removing redundant parameters. Quantisation can also reduce the memory usage of AI models by reducing the bit-width of the model parameters. These techniques not only reduce the size of AI models, but also can significantly decrease the computation and memory usage. In certain circumstances, GPUs or TPUs can also be leveraged as hardware accelerators to execute AI-related tasks, providing extra computational speed to handle AI tasks offline. These techniques, along with the development of power-efficient chip design, helped to overcome hardware challenges of IoT devices and enable AI on these devices more effectively [2].

Combining model compression, quantisation and hardware acceleration for Al algorithms can reduce the computational power needed to 20 per cent of its original amount. This is shown by the formula:

$$\mathrm{P}' = \frac{\mathrm{P}\times(\mathrm{n}\times\mathrm{C}\times\mathrm{Q})}{\mathrm{H}} \tag{1}$$

This formula quantifies the reduction in computational power (P') based on the original power requirement (P), the number of parameters (n),the compression ratio (C), the quantization factor (Q), and the hardware acceleraton efficiency (H).

2.2. Software Architectures for AI-Driven IoT

The integration of AI into IoT systems also requires a reconfiguration of traditional software architectures to address the specific needs of AI-driven processing. Decentralised (or fog) computing architectures have emerged as critical enablers of AI in IoT because they allow data to be processed closer to the physical artefact rather than at faraway clouds or data centres. When data needs to be processed in real time, as is the case in critical IoT applications like smart grids or autonomous vehicle safety, using decentralised AI architectures that distribute processing across multiple nodes in the network allows for higher efficiency of data processing and decisionmaking.Edge computing is similar in principle to decentralised computing, with the notable difference that processing happens at the edge of the network rather than within an on-premise data centre. It too allows for computation and data processing to happen closer to the physical artefact rather than at faraway clouds, reducing the need for data to traverse extensive distances, minimising latency and improving response times [3]. The software architectures that support AI-driven IoT must allow for scalability and accommodate AI requirements such as real-time data processing, fault tolerance and dynamic resource allocation.

2.3. Data Management and AI Integration

Data management also plays a crucial role. The performance of AI algorithms is dependent on the quality, quantity, and timeliness of the data managed by the system. Data in IoT environments is often generated in large volumes and high speeds, making it challenging for storage and processing at the right time. Effective data management is necessary to ensure that the AI algorithms are fed with the right data when they are supposed to be, and that they are fed with data of high quality so that the algorithms can come to the right decision. A key data management challenge to information collected from multiple devices or from external sources is the sheer amount of distributed data that needs to be managed. Distributed storage, for example in the form of a blockchain, helps to deal with this challenge in terms of data management. Data needs to be stored in a distributed and decentralised format to secure integrity and immutability of data, and making it accessible to the AI algorithm to create a learning mechanism [4]. Another management challenge related to data is that of preprocessing data. Noise and redundant/irrelevant information has to be filtered before the data is fed to the AI algorithms. Such data may be generated in high velocity in IoT environments.

3. Optimization Techniques for AI in IoT

3.1. Energy-Efficient AI Algorithms

Energy efficiency is another important issue as many IoT devices are battery-powered and need to operate for months or even years without recharging the battery. The energy-expensive nature of traditional AI algorithms poses a challenge for deploying AI to IoT devices; thus, there is a need to develop energy-efficient AI alternatives. The energy efficiency of AI algorithms can be improved by designing algorithms that require less computation power, which in turn minimises the power consumption. Model pruning is a technique that keeps the important parameters of the AI model and removes the rest. Specific AI model designs for low power devices are another trick to reduce the computational load. Another option is to optimise the execution of AI algorithms on hardware. For example, AI algorithms can be executed in low-power processing modes of the hardware. Task scheduling can also be used to distribute the computational load on the hardware in periods of low energy consumption. Energy harvesting where energy from the environment, like solar power, is captured and stored to meet the energy demand of IoT devices is another technique to improve energy efficiency in IoT [5]. Table 1 includes numerical data for various energy-efficient AI algorithms in IoT environments.

Table 1. Impact of Energy-Efficient AI Techniques on Power Consumption and Battery Life in IoT Devices

Energy-Efficient Technique	Power Consumption Reduction (%)	Battery Life Extension (%)	Example Implementation (Power Saved per Day in mAh)
Model Pruning	30	40	200
Lightweight AI Models	25	35	150
Low-Power Processing Modes	20	30	120
Task Scheduling	15	20	100
Energy Harvesting	10	50	80

3.2. Latency Reduction in AI-Driven IoT Systems

Latency is one of the critical performance issues in AI-driven IoT systems, particularly in applications that require decisions to be made in real-time, such as autonomous vehicles or in industrial automation. If the latency is high, it could lead to delays in processing data and making decisions, which could have serious consequences in many applications. There are many techniques that have been devised to reduce latency in the AI-driven IoT system. One approach is to perform processing at the edge of the network, closer to where the data is generated, rather than sending the data to a central data centre for processing. This can help to reduce the time taken for data to be processed and decisions to be made, as well as to reduce the amount of data that needs to be transmitted over the network. Another approach is to use AI accelerators, hardware specifically designed to perform AI tasks more efficiently than general-purpose processors. AI accelerators can significantly reduce the time needed to run the AI algorithms, thus reducing the latency [6]. Another technique to further reduce latency is to employ pipelining and parallel processing, which allow multiple tasks to be processed at the same time. In this way, we can minimise latency in the AI-driven IoT system so that decisions can be made in real-time.

3.3. Scalability of AI-Driven IoT Systems

With the rising usage of IoT devices, scalability will become an important aspect of AI-driven IoT systems. Scalability is the ability of a system to handle an increasing number of things, such as nodes, devices, data or processes while delivering the same level of service. Scalability challenges of AI-driven IoT systems derive from the ability of AI algorithms to process an increasing amount of data generated from a huge network of devices. To overcome these challenges, AI algorithms can be executed on distributed AI frameworks, where the execution of AI operations is distributed across multiple devices or nodes of the network [7]. Distributed AI frameworks reduce the burden on each device and allow for increasing the number of devices or data streams that the system can handle. Cloud platforms are also fundamental pillars of the scalability of AI-driven IoT systems, using the massive computing resources and storage capacity of the cloud to process large amounts of data. Cloud-based platforms can scale in terms of the number of devices or data streams, allowing AI-driven IoT systems to grow at a pace that is much faster than the one of traditional systems, without sacrificing the level of service delivery. Furthermore, the use of load balancing, dynamic resource allocation and similar techniques can enhance the scalability of AI-driven IoT systems by ensuring the efficient use of resources [8]. Figure 1 shows how the impact of AI-driven IoT systems changes when different scalability techniques are applied.

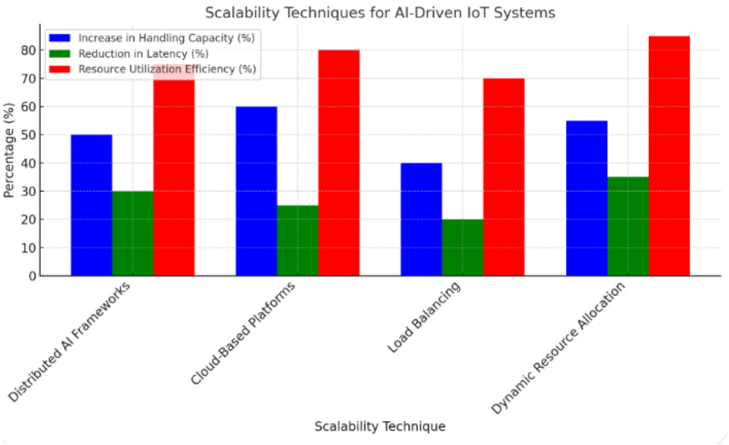

Figure 1. Scalability Techniques for AI-Driven IoT Systems

4. Real-Time Data Processing in AI-Driven IoT Systems

4.1. Real-Time Data Analytics

Real-time data analytics is another key aspect of AI-driven IoT systems, as they need to process and analyse data as it is created, and take decisions in real-time. Optimising AI for real-time data analytics requires us to tackle several key challenges. First, we need to process real-time data streams efficiently and in a timely manner, so that it doesn't introduce substantive latency. This can be done by using streaming data processing techniques to process the data in small chunks as it arrives, instead of waiting to collect the full dataset. Additionally, the AI algorithms used in real-time analytics should be able to handle the dynamic nature of IoT data, which can be highly variable in volume, velocity and quality. This requires adaptive AI algorithms that can adjust the processing strategy based on the characteristics of the data [9]. Finally, real-time data analytics in AI-driven IoT systems should also be scalable, as the number of data streams and devices can grow rapidly. Optimising AI for real-time data processing and being able to handle the unique challenges of IoT environments can enable real-time decision making in AI-driven IoT systems.

4.2. Edge AI for Real-Time Decision-Making

Edge AI refers to the execution of AI algorithms at the edge of the network, very close to where data is generated. This can help ensure real-time decision-making in systems of IoT that reduce human interaction and control. The key advantage of edge AI is that it allows processing data on-site, in systems such as autonomous vehicles, industrial automation and smart healthcare, where response times must be near-instant due to the critical nature of decisions. To implement edge AI, so-called lightweight AI models, specifically designed to be executed in edge environments, are required. As with fog and cloud AI, lightweight models must be adapted to restricted computational resources, such as those available on edge devices, to execute in real time and with low latency. Additionally, more effort is required towards developing adaptive AI algorithms capable of processing data in an ever-changing environment, as data produced at the edge is inherently dynamic, possibly unstable and changeable in real time [10]. This requires that edge AI systems must be able to dynamically adjust their processing strategies to the current state of data. Furthermore, considering the dynamics of IoT that increases exponentially the number of devices and data streams at the edge of the network, edge AI systems must also be scalable. In this scenario, the development of lightweight, adaptive AI models that can operate in edge environments is the key technology enabler for the implementation of real-time decision-making in AI-driven IoT systems [11]. The examples in table 2 show specific applications for edge AI, how lightweight AI models are utilised to achieve significant latency reduction, scalability and real-time decision-making capabilities in multiple industry areas.

Table 2. Edge AI Applications for Real-Time Decision-Making

Edge AI Application	Lightweight AI Model Used	Latency Reduction (%)	Scalability Factor	Real-Time Decision-Making Capability
Autonomous Vehicles	Optimized Convolutional Neural Network (CNN)	85	High - Can manage thousands of vehicles	Critical - Millisecond-level decision-making
Industrial Automation	Real-Time Predictive Maintenance Model	75	Moderate - Scales to large factories	High - Near real-time equipment monitoring
Smart Healthcare	Low-Power Diagnostic AI Model	80	Moderate - Scales to multiple healthcare devices	Critical - Instantaneous patient monitoring
Smart Cities	Decentralized Traffic Management AI	70	High - Scales across city-wide infrastructure	Moderate - Real-time traffic adjustments
Agricultural Monitoring	Crop Growth Prediction AI	65	Low - Scales to regional farms	Moderate - Hourly crop status updates

5. Conclusion

Specifically, this paper addresses AI-integration related challenges in IoT systems, with an emphasis on resource-constrained settings. It answers this question in two ways. First, it illustrates how advanced optimisation techniques, such as model compression or the use of hardware accelerators, enables AI algorithms to operate in a resource-constrained environment, such as IoT devices. Second, it shows that decentralised and edge computing architectures can help alleviate the latency and increase the scalability of AI-based IoT systems. Third, it demonstrates the importance of energy-efficient AI algorithms, which can prolong the usable life of battery-powered IoT devices. Finally, this paper provides practical insights into the technical challenges faced by AI-driven IoT systems and proposes a set of solutions, which can serve as a guide for future developments. This allows these systems to continue to evolve and respond to the increasing demands of processing and making individualised decisions in real-time across multiple domains.

References

[1] Saied, Mohamed, Shawkat Guirguis, and Magda Madbouly. "Review of artificial intelligence for enhancing intrusion detection in the internet of things." Engineering Applications of Artificial Intelligence 127 (2024): 107231.

[2] Charef, Nadia, et al. "Artificial intelligence implication on energy sustainability in Internet of Things: A survey." Information Processing & Management 60.2 (2023): 103212.

[3] Purnama, Suryari, and Wahyu Sejati. "Internet of things, big data, and artificial intelligence in the food and agriculture sector." International Transactions on Artificial Intelligence 1.2 (2023): 156-174.

[4] Hadzovic, Suada, Sasa Mrdovic, and Milutin Radonjic. "A path towards an internet of things and artificial intelligence regulatory framework." IEEE Communications Magazine 61.7 (2023): 90-96.

[5] Ubina, Naomi A., et al. "Digital twin-based intelligent fish farming with Artificial Intelligence Internet of Things (AIoT)." Smart Agricultural Technology 5 (2023): 100285.

[6] Soori, Mohsen, Behrooz Arezoo, and Roza Dastres. "Internet of things for smart factories in industry 4.0, a review." Internet of Things and Cyber-Physical Systems 3 (2023): 192-204.

[7] Andronie, Mihai, et al. "Big data management algorithms in artificial Internet of Things-based fintech." Oeconomia Copernicana 14.3 (2023): 769-793.

[8] Abed, Ali Kamil, and Angesh Anupam. "Review of security issues in Internet of Things and artificial intelligence-driven solutions." Security and Privacy 6.3 (2023): e285.

[9] Tomazzoli, Claudio, Simone Scannapieco, and Matteo Cristani. "Internet of things and artificial intelligence enable energy efficiency." Journal of Ambient Intelligence and Humanized Computing 14.5 (2023): 4933-4954.

[10] Zaidi, Abdelhamid, et al. "Smart Implementation of Industrial Internet of Things using Embedded Mechatronic System." IEEE Embedded Systems Letters (2023).

[11] Tyagi, Amit Kumar. "Blockchain and Artificial Intelligence for Cyber Security in the Era of Internet of Things and Industrial Internet of Things Applications." AI and Blockchain Applications in Industrial Robotics. IGI Global, 2024. 171-199.