21. GI/ITG KuVS Fachgespräch Sensornetze (FGSN 2024), Osnabrück, Germany, September 2-3, 2024

Challenges in Developing Modular AI Applications

Kamil Wołoszyn, Krzysztof Piotrowski, Krzysztof Turchan
IHP - Leibniz Institut für innovative Mikroelektronik
Frankfurt (Oder), Germany
{woloszyn, piotrowski, turchan}@ihp-microelectronics.com

***Abstract*—The development of modular structures for AI (Artificial Intelligence) applications addresses the growing complexity and demand in this field. Modular AI applications, designed for rapid adaptation and deployment in diverse environments, offer significant advantages alongside notable challenges. This paper discusses these challenges and introduces concepts such as AI4U, Sens4U, and AMAD, which aim to provide effective solutions. Key issues include achieving modularity, optimizing for resource-constrained edge devices, automating sensor configurations, and ensuring robust data privacy and security measures. By leveraging these frameworks, the paper demonstrates how to enhance the efficiency, scalability, and security of AI applications. Additionally, the paper highlights various application areas—such as smart cities, agriculture, surveillance, and environmental monitoring—illustrating the broad potential and adaptability of modular AI systems.**

***Index Terms*—modular, automating, concepts, AI**

I. Introduction

The increasing complexity and demand for AI applications have forced the development of modular structures that allow for flexibility, scalability, and efficiency. Modular AI applications that can be rapidly developed and adapted to various environments offer significant benefits but also come with substantial challenges. This work presents developed concepts and problems that must be addressed when creating modular AI applications. The concept of modular AI applications involves creating systems that can integrate various sensors and modules to efficiently perform specific tasks. Concepts such as AI4U [1] and AMAD [2] have been developed based on the Sens4U [3] approach to facilitate the entire application creation process. These structures emphasize modularity, allowing for easy addition, removal, or replacement of different components without significant changes to the entire system. Creating such applications involves a range of technical and organizational problems. The main challenges involve dealing with computational and memory resource limitations on edge devices, which often have significantly lower capabilities than central devices. Optimizing AI models is crucial to ensure their efficient operation on these devices. This paper aims to discuss both theoretical and practical aspects of developing modular AI applications. Both developed concepts and specific problems and requirements that developers must face in this dynamically developing technology area will be presented.

II. Challenges and Proposed Solutions in Developing Modular AI Applications

A. Modularity and Interoperability

Ensuring that different modules can work together seamlessly is a significant challenge. This requires standard interfaces and protocols that can handle communication between various components regardless of their origin. AI4U addresses this problem by defining interfaces and utilizing a modular architecture (see Fig. 1), that supports interoperability. AI4U

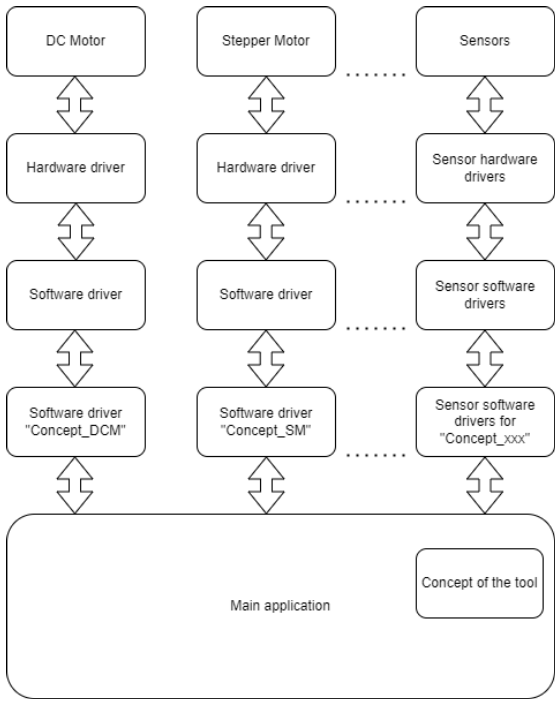

Fig. 1. Concept of the AI4U tool

is a concept of modular structures aimed at accelerating the creation of artificial intelligence applications. The main goal of AI4U is to ensure versatility and flexibility, allowing easy adaptation of AI applications to various environments. These structures integrate different types of sensors, enabling the efficient performance of specific tasks. One of the key assumptions of AI4U is modularity, which allows for rapid application development through the easy addition, removal, or replacement of components without modifying the entire system. One of the main goals of AI4U is to enable the creation of AI applications even by non-experts in technology.

For this purpose, the structure uses modular libraries that facilitate building and testing AI applications. AI4U is designed with a wide range of applications in mind, from simple consumer applications to more advanced research solutions. The entire AI4U concept is based on the Sens4U approach, which introduces fundamental principles of modularity and interoperability into the AI4U structure. Sens4U focuses on using modular libraries and standard interfaces, allowing for quick and efficient integration of various sensors and components to perform complex tasks. Thanks to this, AI4U inherits the ability to easily add, remove, and replace modules, which is crucial for creating flexible and scalable AI applications.

B. Resource Constraints on Edge Devices

Edge devices, often used in modular AI applications, have limited computational and memory resources. Optimizing AI models and algorithms to work efficiently on these devices is crucial. Using frameworks such as TensorFlow Lite [4] and PyTorch Mobile [5], which offer optimized versions of deep learning libraries, is essential to overcoming these limitations. For example, the TensorFlow Lite framework allows running models on devices with limited computational power, which is crucial for applications running on edge devices. Optimizing AI models for edge devices involves several key aspects. First, the model size must be reduced to fit within the device's limited memory. Techniques such as quantization, which reduces the precision of numbers used in the model, and pruning, which removes unnecessary connections in the neural network, are often used. These techniques can significantly reduce the model size without significantly impacting its accuracy. The second aspect is optimizing for energy consumption. Edge devices, especially those running on batteries, must be energy-efficient. Therefore, algorithms must be designed to use as little energy as possible. Techniques such as dynamic voltage and frequency scaling (DVFS) [6] can be used to reduce the processor's energy consumption during less intensive operations. Another challenge is ensuring low latency. AI applications on edge devices often need to operate in real-time, which means that data processing delays must be minimal. Latency optimization can involve both adjusting the model architecture and implementing algorithms to enable parallel data processing. In addition to TensorFlow Lite and PyTorch Mobile, there are other tools and frameworks that support AI model optimization for edge devices. For example, ONNX Runtime [7] offers support for running machine learning models on various hardware platforms, optimizing their performance. Another example is Google's Edge TPU [7], which provides hardware support for accelerating AI model inference on edge devices.

C. Automated Sensor and System Configuration

Selecting and configuring the appropriate sensors and system components is another challenge in creating modular AI applications. The AMAD (Automated Modular Application Development) concept automates this process by using metadata to describe sensor characteristics, such as energy consumption and interface specifications, allowing for precise selection and configuration (see Fig. 2). One of the key

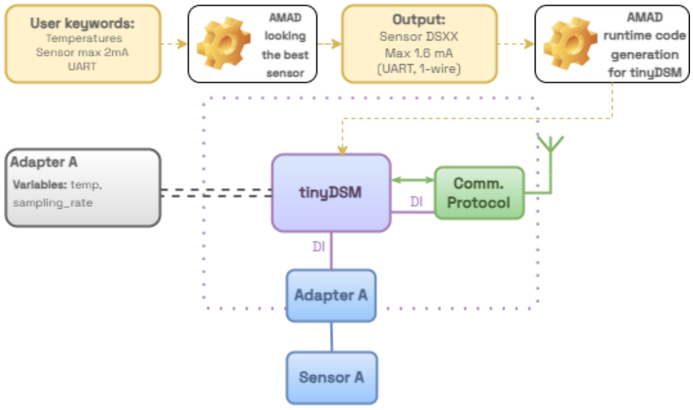

Fig. 2. Flow of the AMAD tool

features of AMAD its user-friendly interface, which allows users to input specific metadata regarding their requirements. Users can specify parameters such as what they want to measure, the communication protocol they prefer to use, and the maximum current they want the sensor to consume during measurements. Based on these inputs, the AMAD tool processes the information and provides a recommendation for the best sensor available in its database that meets the user's criteria. This automated system helps users quickly find the most suitable sensor, its connection and the code to operate it without requiring extensive technical knowledge. This makes the configuration process faster and less prone to human errors, which is crucial in dynamic work environments. The AMAD tool concept facilitates the management of complex systems by automatically adjusting sensors to specific tasks and operating conditions. Using metadata allows for more efficient and precise resource management, which is particularly important given the limited computational and energy resources of edge devices. Thanks to AMAD, systems can be quickly reconfigured and optimized without the need for manual intervention, significantly speeding up the deployment and adaptation process of AI applications to new conditions.

D. Data Privacy and Security

Processing data locally on edge devices can significantly enhance data privacy and security by reducing the need to transmit sensitive information to remote servers. Local data processing limits the risk of data interception during transmission and allows for greater control over information. However, to fully utilize these benefits, solid security measures must be implemented to protect data on the device itself. Ensuring data confidentiality requires the implementation of advanced encryption mechanisms that secure data both at rest and in transit. However, local data processing on edge devices also brings challenges. These devices must be regularly updated to protect them against the latest security threats. Additionally, monitoring systems must be implemented to detect and respond to potential security incidents in real-time.

E. Application Areas

Modular AI applications have a wide range of potential applications, including smart cities, agriculture, surveillance and security, and environmental monitoring. Each of these areas benefits from the flexibility and scalability offered by modular AI structures. In smart cities, AI applications can be used to manage traffic, monitor the environment, and optimize energy consumption. For example, AI-based traffic management systems can analyze real-time data from various sensors to optimize traffic signals and reduce congestion. In agriculture, modular AI applications can significantly improve efficiency and sustainability. AI systems can monitor soil conditions, humidity, temperature, and other environmental factors to provide precise information on irrigation and fertilization. In the industrial sector, modular AI applications can optimize production processes, monitor machine conditions, and predict failures. AI systems can analyze data from sensors placed on production lines to identify potential problems.

III. Conclusions

Creating modular AI applications involves many challenges, but existing concepts such as AI4U, Sens4U, and AMAD present solutions that help address these problems. These structures provide advanced tools and methodologies that enable more efficient and scalable AI application development. One of the most significant challenges in creating modular AI applications is ensuring modularity. Modularity allows for easy addition, removal, or replacement of individual components without the need to modify the entire system. AI4U and Sens4U demonstrate how to minimize this problem by defining standard interfaces and communication protocols that enable seamless integration of different modules. This allows systems to be quickly adapted to changing needs and operating conditions, which is crucial in dynamic work environments. Conceptual approaches such as AI4U, Sens4U and AMAD provide example solutions to key issues of modularity, resource constraints, automated configuration and data security. These approaches not only accelerate the AI application development process but also ensure that they are more flexible, scalable, and secure. Thanks to them, it is possible to utilize the full potential of artificial intelligence in various fields, contributing to technological development.

References

[1] K. Wołoszyn, K. Turchan, M. Rąpała, K. Piotrowski, AI4U: Modular Framework for AI Application Design

[2] K. Wołoszyn, K. Piotrowski, Automated Modular Application Development, in Proc. 15th Konferencja Naukowa Systemy Pomiarowe w badaniach naukowych i w przemyśle (SP 2024), 2024

[3] Piotrowski K., Peter St.: Sens4U: Wireless Sensor Network Applications for Environment Monitoring Made Easy, in Proc. 4th International Workshop on Software Engineering for Sensor Network Applications (SESENA 2013), 2013.

[4] TensorFlow Lite, https://www.tensorflow.org/lite, last viewed 16.07.2024

[5] PyTorch Mobile, https://pytorch.org/mobile/home/, last viewed 16.07.2024

[6] Dynamic voltage and frequency scaling, https://www.techtarget.com/whatis/definition/dynamic-voltage-and-frequency-scaling-DVFS, last viewed 16.07.2024

[7] ONNX Runtime, https://onnxruntime.ai/, last viewed 16.07.2024