EMELIOT: Engineered MachinE Learning-intensive IoT systems

Luciano Baresi
Polytechnic University of Milan
Milan, Italy
luciano.baresi@polimi.it

Andrea De Lucia
University of Salerno
Fisciano, Italy
adelucia@unisa.it

Massimiliano Di Penta
University of Sannio
Benevento, Italy
dipenta@unisannio.it

Davide Di Ruscio
University of L'Aquila
L'Aquila, Italy
davide.diruscio@univaq.it

Leonardo Mariani
University of Milano - Bicocca
Milan, Italy
leonardo.mariani@unimib.it

***Abstract*—The rapid growth of Internet-of-Things (IoT) devices and the increasing amount of data they generate pose challenges for traditional cloud-based analysis. The need for distributed and federated solutions has emerged, where data analysis is performed cooperatively on the devices. While the research community has focused on developing distributed machine learning (ML) algorithms, creating dependable ML-intensive IoT systems remains an open challenge.**

EMELIOT (Engineered MachinE Learning-intensive IoT systems) is a National Italian PRIN project initiated in 2022 to seek innovative solutions to various challenges associated with ML-based IoT systems, including multi-task and decentralized learning, scalability, reliability, data security, and performance. The project strives to advance the IoT ecosystem through its interdisciplinary approach and facilitates the widespread adoption of ML technologies in IoT applications.

I. Motivation and Context

Forecasts say that in 2025 there will be more than 25 billion Internet-of-Things (IoT) devices connected to the web [1], while the combination of more devices and more data already created a problem: the well-known solution where all data are transferred onto the cloud and analyzed there as a whole does not pay off anymore [2]. The size of generated data, the high execution times and delays (latency), and the need for data privacy- and ownership-preserving solutions call for distributed [3] and federated solutions [4], where training and inference activities are carried out cooperatively on the different devices.

While the research community has mainly focused on developing Machine Learning (ML) algorithms that are suitable for distributed settings [5], recent experience showed that developing highly dependable ML-intensive IoT systems is to a large extent an open challenge, as also discussed in multiple surveys [3], [6]. In particular, methods and techniques that support multi-task and decentralized learning, scalability, reliability, data security, and performance are still under-explored and significant improvements are needed. As concluded in the survey by Verbraeken et al. [3], more studies are needed for the development of production-level systems.

EMELIOT (Engineered MachinE Learning-intensive IoT systems) is a National Italian project started in 2022 that aims to integrate approaches from diverse areas of software engineering to develop innovative solutions to tackle most of the aforementioned problems and deliver methods and tools for the creation of production-ready ML-based IoT systems. More in detail, EMELIOT wants to find novel solutions to the following challenges.

High heterogeneity: Designing and developing an ML-Intensive IoT system requires dealing with hardware components, ML algorithms, software modules, cloud services, and network protocols whose integration poses development problems not encountered in any other domain.

Multi-disciplinary context: Engineering and operating a ML-intensive IoT system is intrinsically challenging since it requires a broad variety of (sometimes very specific) pieces of knowledge that are not always available inside teams, including knowledge about software engineering practices, machine learning algorithms, hardware components, distributed systems, and network protocols. *Continuously Learning Components*: IoT systems exploit ML modules, that are continuously learning components whose behavior can neither be fully predicted nor verified once for all. Ensuring the dependability of the system is thus an activity that cannot stop with the development phase, but must necessarily continue in the field, while the system is operating.

Efficiency of the employed ML systems: Data processing services should be prompt and reactive to user requests. However, when data are analyzed mainly by centralized solutions, the performance of the whole system might be affected due to the unavoidable time needed to transmit the huge amount of data from the related sources to the remote cloud. This calls for distributed environments that can support scalable and efficient model training and usage.

End-to-end Reliability: The paradigm shift induced by the adoption of ML components and their integration within IoT systems makes reasoning about the behavior of software systems difficult since they are intrinsically challenging to test and verify. Indeed, it is hard to exercise scenarios that involve many sensors and actuators controlled by remote ML-based

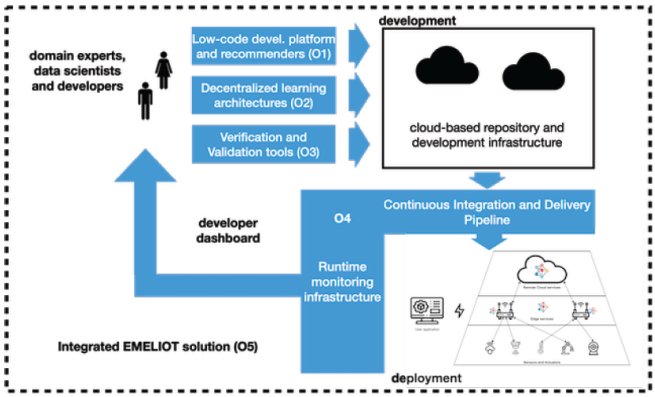

Fig. 1. The EMELIOT Approach.

services, and it is even harder to specify oracles that can accurately capture misbehaviors.

Extended attack surface: The attack surface of a ML-Intensive IoT system is particularly large. On one hand, ML-intensive IoT systems can be subject to attacks leveraging their high distribution *Lack of development and operation processes tailored for ML-based IoT systems*: ML-based IoT systems make use of bespoke ML components, which have limited reuse capabilities. In particular, generic ML packages are merged with the final system by means of glue code, which makes the detection of errors and their repair difficult and costly.

The above challenges are addressed by exploiting the unique combination of expertise of the five universities involved in the project: Polytechnic University of Milan, University of Salerno, University of Sannio, University of L'Aquila, and University of Milano-Bicocca.

II. The EMELIOT Approach

EMELIOT has the goal of providing a comprehensive set of methodologies, architectures, and tools to better develop, verify, and operate ML-intensive IoT systems. EMELIOT targets interdisciplinary teams consisting of different figures including software engineers, data scientists, and ML experts, reducing the distances among such specialists and facilitating their collaboration. The EMELIOT approach combines the technical objectives described below as illustrated in Figure 1.

Development democracy: EMELIOT aims to democratize the development of ML-intensive IoT systems, also supporting non-expert developers in the creation of ML-intensive IoT systems. To this end, EMELIOT will design languages for modeling ML-Intensive IoT systems and will provide users with constructs to specify the peculiar elements characterizing these systems, such as data sources, data analysis tasks, and machine learning techniques to be employed.

Support to Distributed Learning: EMELIOT aims to provide methodologies, tools, and architectures that support and simplify the adoption of federated and distributed learning mechanisms for ML-intensive IoT systems. Modern computing systems are more and more a continuum that starts from IoT and mobile devices and then evolves through edge and fog nodes and often involves cloud-based elements. EMELIOT exploits this continuum to deliver ML-based solutions that can use it seamlessly.

V& V solutions for ML-Intensive IoT Systems: EMELIOT aims to develop verification and validation techniques, including techniques to mitigate security threats, specifically tailored towards ML-intensive IoT systems. EMELIOT tackles key challenges related to the analysis and testing of ML-intensive IoT systems. On one hand, IoT systems rely on non-conventional interfaces, based on a combination of sensors (e.g., motion, temperature, and light sensors) and natural language interfaces (e.g., assistants and chatbots accepting voice and text commands), through which users can interact and that are particularly hard to verify for correctness and security. On the other hand, the lack of oracles further challenges establishing the correctness of the tested behaviors. Finally, non-functional testing, and security testing, in particular, must consider that IoT systems often treat confidential information or sensitive data that may be maliciously manipulated, poisoned, or even stolen ML components that expose IoT systems to security risks that must be timely identified.

CI/CD for ML-Intensive IoT Systems: EMELIOT aims to provide solutions for the continuous integration, delivery, and monitoring of ML-intensive IoT systems. EMELIOT aids developers to operate on and monitor ML-intensive IoT system pipelines. This involves the definition of novel MLOps strategies and techniques, new monitoring systems that data scientists and developers can exploit to continuously verify the inner working of the deployed systems and diagnose potential threats in advance.

Integrated MLOps Infrastructure: EMELIOT aims to ultimately integrate the developed solutions into an infrastructure that can support the whole life-cycle development of ML-Intensive IoT Systems.

References

[1] Statista, "Internet of things (iot) and non-iot active device connections worldwide from 2010 to 2025," Statista. https://www.statista.com/statistics/1101442/iot-number-of-connected-devices-worldwide/.

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, pp. 1–25, 06 2019.

[3] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. Rellermeyer, "A survey on distributed machine learning," *ACM Comput. Surv.*, vol. 53, no. 2, mar 2020.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, jan 2019.

[5] P. Kairouz and H. McMahan, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1, 2021.

[6] S. Lo, Q. Lu, C. Wang, H.-Y. Paik, and L. Zhu, "A systematic literature review on federated machine learning: From a software engineering perspective," *ACM Comput. Surv.*, vol. 54, no. 5, may 2021.