Popular science summary of the PhD thesis

PhD student	Alessandro Cerioli
Title of the PhD thesis	Neural Network Optimization for Predictable Real-Time Edge Devices
PhD school/Department	Department of Applied Mathematics and Computer Science

Science summary

* Please give a short popular summary in English (approximately half a page) suited for the publication of the title, main content, results and innovations of the PhD thesis also including prospective utilizations hereof. The summary should be written for the general public interested in science and technology.

Artificial intelligence is becoming part of everyday life, integrated into applications such as speech recognition, image processing, and smart assistants. However, many of these technologies are computationally demanding and rely on powerful cloud systems. Running artificial intelligence directly on small, low-power devices such as wearable gadgets, industrial sensors, or autonomous embedded systems remains a major challenge.

This thesis investigates how to make neural networks more efficient and suitable for real-time operation on resource-constrained devices. Such systems often have strict limitations in terms of computing power, memory, and energy consumption, and must frequently meet precise timing requirements. These constraints make the direct deployment of modern AI models difficult or even impractical.

To address these issues, this work proposes new techniques for optimizing neural networks so that they can run efficiently on embedded hardware. A key contribution is the development of NeuralCasting, a compiler that transforms neural network models into portable C code, enabling execution without relying on complex runtime environments. This approach improves portability, reduces memory overhead, and enhances predictability in execution.

The research further explores the use of quantization, a technique that reduces the numerical precision of neural networks to decrease computational cost while maintaining accuracy. In particular, mixed-precision quantization strategies are studied to balance performance and energy efficiency. The thesis also investigates the integration of optimized neural networks with time-predictable computing architectures, enabling reliable execution in real-time applications such as speech enhancement.

Overall, this work contributes to advancing the deployment of artificial intelligence at the edge, making AI systems more efficient, predictable, and accessible for use in embedded and real-time environments.

Please submit the summary to the department PhD coordinator together with your thesis