Politecnico di Torino
1859

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

COMPUTER AND CONTROL ENGINEERING

MUR DM 118 - Advanced Deep Learning Optimization for Extreme Edge Applications

Funded By	MINISTERO DELL'UNIVERSITA' E DELLA RICERCA [P.iva/CF:97429780584] Politecnico di TORINO [P.iva/CF:00518460019]

Supervisor	JAHIER PAGLIARI DANIELE - daniele.jahier@polito.it

Contact	MACII ENRICO - enrico.macii@polito.it JAHIER PAGLIARI DANIELE - daniele.jahier@polito.it

Context of the research activity	The excellent accuracy of deep learning models comes at the cost of high complexity. Optimizing these models, reducing their latency, memory, and energy costs, is paramount to enable their execution at the edge. Going in this direction, this project explores: 1. Neural Architecture Search (NAS) to optimize the network topology 2. Self-supervised learning to increase the accuracy of small models while not requiring enormous datasets, 3. Continual learning to cope with data and concept drift. Progetto finanziato nell'ambito del PNRR – DM 118/2023 - CUP E14D23001770006

	Deep learning models are often deployed on edge devices for various tasks, ranging from biosignals processing to industrial assets monitoring. However, these models can have high computational complexity, which is an obstacle for their deployment on low-power edge devices, with limited resources in terms of memory, energy, and processing power. In order to achieve the required accuracy at the minimum latency, it is crucial to optimize deep learning models for these constraints. The candidate will explore three main directions to maintain/improve accuracy while reducing the complexity of deep learning models for real edge applications. The three techniques will be developed in parallel and in synergy with each other, and the work will be done in collaboration with EDA group researchers already working on those topics. The first investigated approach will be Neural Architecture Search (NAS). These techniques automate the process of designing neural network architectures by using optimization algorithms to search through the space of possible architectures and select the ones that perform the best on a given task. Using NAS makes it possible to find network topologies that are

Objectives

simultaneously accurate and efficient in terms of latency and energy consumption, making them more suitable for deployment on edge devices.

The candidate's goal will be to explore different NAS algorithms and evaluate their performance on various edge-relevant tasks. This could involve implementing and comparing different NAS algorithms, as well as experimenting with different search spaces and optimization objectives.

An orthogonal approach to optimize deep learning models for edge devices is to use Self-Supervised Learning (SSL) techniques. SSL is a type of machine learning where the model is trained on data that is not labeled, but rather is generated from some known structure or process. By using SSL, it is possible to increase the accuracy of small models without requiring enormous, labeled datasets, which can be particularly useful since differently from “mainstream” Computer Vision and Natural Language Processing applications, many edge-relevant tasks cannot rely on huge publicly available datasets.
The goal of this second activity will be therefore to improve the accuracy of already optimized models, e.g., those found by NAS algorithms described above. The candidate will apply existing SSL techniques on new tasks, possibly fusing them with NAS.

The final topic considered in this thesis will be Continual Learning (CL), a set of methods that enable a model to adapt to new data over time without forgetting what it has learned previously. This is important because the data distribution on edge relevant tasks can often change over time due to factors such as changes in the environment or in the monitored asset. By using CL techniques, it is possible to cope with this degradation and maintain the accuracy of the model over time. The final goal of this activity will be to deploy an end-to-end working application that exploits previously optimized architectures, and uses CL to maintain high accuracy on a target task over time.

The work plan of the project will be structured as follows:

Months 1-9: Conduct a literature review on: i) NAS algorithms, ii) SLL methods and iii) CL methods, with a focus on edge devices and applications. In particular, the candidate will consider the standard MLPerf Tiny Inference Benchmarks, plus some additional benchmarks related to EDA group projects, focusing on industrial applications, energy management, and biosignals processing.

Months 9-18: Implement and test state-of-the-art NAS, SSL and CL algorithms on the relevant benchmarks described above.

Months 18-36: Deploy and evaluate new NAS algorithms, comparing them with the state-of-the-art in terms of accuracy, latency, and energy consumption of the resulting models. Try to fuse SSL approaches within the developed NASes. Lastly, apply CL techniques on model architectures optimized with NAS + SSL.

Overall, these three approaches - NAS, SSL and CL - offer promising ways to automate the optimization and deployment of deep learning models at the edge and increase their robustness in real-world scenarios. By combining these techniques, the candidate will try to maximize the accuracy of the models, while minimizing its resource requirements in terms of latency, energy consumption, and memory occupation, enabling new tasks to be solved at the edge of the network, without resorting to a centralized approach, or improving the capabilities of existing ones.

Possible publication venues for this thesis include:
- IEEE Transactions on CAD
- IEEE Transactions on Computers
- IEEE Journal on Internet of Things
- IEEE Transactions on Emerging Topics in Computing
- IEEE Transactions on Sustainable Computing
- IEEE Transactions on Neural Networks and Learning Systems
- ACM Transactions on Embedded Computing Systems
- ACM Transactions of Design Automation of Electronic Systems
- ACM Transactions on IoT

The EDA Group has many active industrial collaborations and funded projects on these topics, including:
- TRISTAN (ECSEL-JU 2023)
- ISOLDE (ECSEL-JU 2023)
- HiFiDELITY (ECSEL-JU 2023)
- StorAIge (ESCEL-JU 2021)
- Etc.

Skills and competencies for the development of the activity

1. Experience with neural network models: The project involves designing and evaluating different neural network architectures, so a good understanding of how different types of layers and connections can impact the performance of a model would be important.

2. Familiarity with programming languages: The project will involve implementing and testing various algorithms and techniques, so experience with programming languages such as Python and C is needed.

3. Familiarity with embedded systems and computer architectures: Understanding the constraints and limitations of edge devices is important in order to optimize the deep learning models for these platforms. Experience with working on projects involving edge devices would be useful.