Deep Learning Algorithms Elaboration for Embedded Systems Implementation

PHD FORUM SUBMISSION

Luigi Capogrosso
Department of Engineering for Innovation Medicine, University of Verona, Italy
luigi.capogrosso@univr.it

***Abstract*—Recent advances in deep learning have brought a step-change in the abilities of machines to solve complex problems like object recognition, person detection, and pose estimation. Although these tasks are of great importance in mobile and embedded devices, particularly in areas such as autonomous driving and video surveillance that require sensing and mission-critical applications, current deep learning solutions generally still require significant computational resources to solve them. As a result, deploying these models on embedded devices can result in extended execution times and the utilization of excessive resources, such as CPU, memory, and power. Without a solution, the hoped-for advances in embedded machines and deep will not arrive. Accordingly, this thesis proposes novel techniques that guarantee a smooth transition of deep learning technology from a scientific environment with virtually unlimited computing resources into embedded systems, using *Split Computing* and *TinyML*.**

***Index Terms*—Deep Neural Networks, Embedded Devices, Split Computing, TinyML**

I. Introduction

In the last decade, Deep Neural Networks (DNNs) achieved state-of-the-art performance in a broad range of problems, spanning from object classification and detection to speech recognition and predictive maintenance. This success comes at a price: the computational requirements of some DNNs preclude their deployment on most resource-constraint devices, such as mobile phones available today; we refer to this scenario as *Local-only Computing* (LC). So, the current approach, usually referred to as *Remote-only Computing* (RC), consists of transferring the data captured by the device to a high-performance machine through a communication network and then sending back the result to the device. As a compromise between LC and RC approaches, the ***Split Computing** (SC)* frameworks propose to divide the DNN model into a "head", deployed on the sensing device, and a "tail" deployed on the remote server.

Starting from the above examples, it is understandable that the design of a distributed deep learning application results in moving into a *three-dimensional design space exploration*. Indeed, a given implementation is determined by the choice of the *computation platform*, the *communication architecture*, and the *DNN*. Whereas the first two dimensions are deterministic, in the sense that a given choice of platform or communication architecture brings to a certain performance, dealing with DNNs introduces uncertainty. A DNN is essentially a statistical classifier with numerous parameters and various architectures. Its timing and effectiveness are not deterministic, and accuracy is commonly used to evaluate its performance. When it comes to distributed architectures and SC strategies, the situation becomes even harder, since different types of split require specific training. So, deciding on a properly distributed architecture hosting DNNs and manipulating diverse SC configurations requires days. As a result, in recent years, research has focused heavily on improving embedded technologies for use in resource-limited environments.

In this regard, Micro-Controller Unit (MCU) based embedded systems have garnered tremendous attention, primarily due to the low power requirement and, secondly due to crucial performance and reliability traits such as safety, security, maintainability, and adaptability. This leads to the need to adopt a new paradigm: ***TinyML***. To be more exact, until recently, carrying out machine learning tasks on a microcontroller was deemed impractical. Consequently, there's been a swift and dynamic evolution in hardware, software, and research.

Therefore, we can easily understand how embedded deep learning will open up a series of possibilities for applications in IoT devices, such as smart buildings and other devices so that they have intelligent functionalities that today are restricted to computers and smartphones. We will see voice interfaces in almost everything in the future. For example, as soon as we can create suitable voice interfaces at a low cost, we will have them on any consumer item, replacing buttons on any device, especially if you think of devices combining audio and video.

II. Achieved Results

With regard to the SC, in [1] we propose a fast procedure to select the best-split location for a generic DNN architecture that, for the first time, is predictive of the accuracy that the system will have once retrained. The method is dubbed **I-Split**, where "I" stands for interpretability. I-Split builds upon the concept of importance (or saliency) of a neuron, which is related to the gradient it possesses with respect to the decision towards the correct class, for the specific input. Importance is exploited with success in the Grad-CAM approach: Grad-CAM creates an input neuron saliency map that indicates which parts of an input image are more important for deciding a specific class. In particular, the Grad-CAM approach has been proved to be strongly dependent on the given trained model on which it runs, while other approaches do not, making it perfectly suited to our purposes.

This work was later extended in [2], in which we propose ***Split-Et-Impera***, a novel and practical framework that *i)* determines the set of the best-split points of a neural network based on deep network interpretability principles without performing a tedious try-and-test approach, *ii)* performs a communication-aware simulation for the rapid evaluation of different neural network rearrangements, and *iii)* suggests the best match between the quality of service requirements of the application and the performance in terms of accuracy and latency time.

On the other hand, regarding TinyML, the first work is in the context of smart buildings and smart cities. Specifically, we saw that the design of low-cost and privacy-aware solutions for recognizing the presence of humans and their activities is becoming of great interest. Existing solutions exploiting wearables and video-based systems have several drawbacks, such as high cost, low usability, poor portability, and privacy-related issues. Consequently, more ubiquitous and accessible solutions, such as WiFi sensing, became the focus of attention. However, at the current state of the art, WiFi sensing is subject to low accuracy and poor generalization. Such issues are partially solved at the cost of complex data preprocessing pipelines. In [3], we present a highly accurate, resource-efficient deep learning-based occupancy detection solution, which is resilient to variations in humidity and temperature. The approach is tested on an extensive benchmark, where people are free to move and the furniture layout does change. In addition, based on a consolidated algorithm of explainable AI, we quantify the importance of the WiFi signal w.r.t. humidity and temperature for the proposed approach.

Furthermore, in [4], we present ***DOHMO***, an embedded computer vision system where multiple sensors, and intelligent cameras, are connected to actuators in order to regulate illumination and doors. The system aims at assisting elderly and impaired people in co-housing scenarios, in accordance with privacy design principles. The paper provides details of two core elements of the system: the first one is the BOX-IO controller, a fully scalable and customizable hardware and software IoT ecosystem that can collect, control, and monitor data, whether indoor or outdoor. The second one is the embedded 3DEverywhere intelligent camera, a device composed of an embedded system that receives input data provided by a 3D/2D camera, analyzes it, and returns the metadata of this analysis. We illustrate how they can be connected and how simple decision mechanisms can be implemented in such a framework. In particular, illumination can be triggered on and off by the detected presence of people, overcoming the limitations of typical sensors, while doors can be opened or closed based on person trajectories in an intelligent manner.

A special focus on smart doors was then carried out in [5]. This work proposes the first position paper related to smart doors, without bells and whistles. We first point out that the problem not only concerns reliability, climate control, safety, and mode of operation. Indeed, a system to predict the intention of people near the door also involves a deeper understanding of the social context of the scene through a complex combined analysis of proxemics and scene reasoning. Furthermore, we conduct an exhaustive literature review about automatic doors, providing a novel system formulation. Finally, we present an analysis of the possible future application of smart doors, a description of the ethical shortcomings, and legislative issues.

Still concerning TinyML, in [6] we present the ***Safe Place*** project. Safe Place is an Italian 3M euro regional industrial/academic project, financed by European funds, created to ensure a multidisciplinary choral reaction to COVID-19 in critical environments such as rest homes and public places. Safe Place consortium was able to understand what is no longer useful in this post-pandemic period, and what instead is potentially attractive for the market. For example, the detection of face masks has little importance, while sanitization does have much. This paper shares such analysis, which emerged through a co-design process of three public Safe Place project demonstrators, involving heterogeneous figures spanning from scientists to lawyers.

III. Thesis Completion

In the near future, we plan to continue to improve on what we have achieved so far. Specifically, concerning I-SPLIT, future works will include further investigation of interpretability methods as a way to extract additional metrics to be used in the generation of the I-SPLIT curve. Regarding *Split-Et-Impera*, instead, we will focus on investigating specific techniques of tensor reconstruction to handle packet losses in a UDP transmission.

Finally, regarding TinyML, another important field of study in order to achieve the objectives of this thesis is *neuromorphic computing*, *i.e.*, the use of Very-Large-Scale Integration (VLSI) systems containing electronic analog circuits to mimic neuro-biological architectures present in the nervous system.

References

[1] F. Cunico, L. Capogrosso, F. Setti, D. Carra, F. Fummi, and M. Cristani, "I-split: Deep network interpretability for split computing," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 2575–2581.

[2] L. Capogrosso, F. Cunico, M. Lora, M. Cristani, F. Fummi, and D. Quaglia, "Split-et-impera: A framework for the design of distributed deep learning applications," in *2023 26th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*. IEEE, 2023, pp. 39–44.

[3] C. Turetta, G. Skenderi, L. Capogrosso, F. Demrozi, P. H. Kindt, A. Masrur, F. Fummi, M. Cristani, and G. Pravadelli, "Towards deep learning-based occupancy detection via wifi sensing in unconstrained environments," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–6.

[4] G. Skenderi, A. Bozzini, L. Capogrosso, E. C. Agrillo, G. Perbellini, F. Fummi, and M. Cristani, "Dohmo: Embedded computer vision in co-housing scenarios," in *2021 Forum on specification & Design Languages (FDL)*. IEEE, 2021, pp. 01–08.

[5] L. Capogrosso, G. Skenderi, F. Girella, F. Fummi, and M. Cristani, "Toward smart doors: A position paper," *arXiv preprint arXiv:2209.11770*, 2022.

[6] F. Cunico, L. Capogrosso, A. Castellini, F. Setti, P. Pluchino, F. Zordan, V. Santus, A. Spagnolli, S. Cordibella, G. Gennari *et al.*, "The post-pandemic effects on iot for safety: The safe place project," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–4.