Advancing Split Computing and Anomaly Detection for On-Edge and Interpretable Industry 4.0

PHD FORUM SUBMISSION

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

***Abstract*—This thesis explores the synergy of Split Computing (SC) and Anomaly Detection (AD), presenting novel methodologies that enhance computational efficiency and detection accuracy within an Industry 4.0 scenario. In particular, we investigate the application of SC frameworks, where a Deep Neural Network (DNN) is intelligently split with a part of it deployed on an edge device and the rest on a remote server, to optimize resource allocation in edge and cloud environments, enabling real-time processing on large-scale industrial data. Additionally, to significantly improve the robustness of defect detection systems, we introduce novel techniques based on generative Artificial Intelligence. Specifically, we enhance Diffusion Models for defect image generation, which significantly improves the robustness of defect detection systems. As a result, this thesis demonstrates a substantial leap in contributing to on-edge, scalable, efficient, and interpretable solutions for AD in Industry 4.0.**

I. Introduction

Four major evolutions of industrialization have occurred throughout human history, impacting economic growth, population expansion, and major social transformations. The most recent of these, known as Industry 4.0, has process automation as its main priority, thereby reducing the intervention of humans in the manufacturing process, and since, in the last decade, Deep Neural Networks (DNNs) achieved state-of-the-art performance in a broad range of problems, aligns perfectly with the goals of Industry 4.0.

A fundamental role in Industry 4.0 is played by Predictive Maintenance (PdM) since it guarantees the ongoing reliability, efficiency, and optimal functionality of advanced technological systems. The first step in PdM is Anomaly Detection (AD), which focuses on detecting abnormal behavior in the equipment by analyzing the historical data of the equipment. However, in the fast-changing world of industrial anomaly detection, there are two main challenges. First, the systems need to be accurate in identifying anomalies. Second, they need to be efficient and use minimal resources, as many edge and industrial devices have limited processing power.

This thesis addresses this need by integrating advancements in the field of Split Computing (SC), where a DNN is intelligently split with a part of it deployed on an edge device and the rest on a remote server for real-time large-scale processing. Moreover, it leverages the latest trends in the Machine Learning (ML) learning research field for defect image generation to significantly improve the robustness of defect detection systems.

II. Split Computing

DNN models often present computational requirements that cannot be met by most of the resource-constraint edge devices available today [1]. This prohibits the full deployment of DNN-based applications on these systems, leading to what is commonly known as the Local-only Computing (LoC) approach. However, using simplified models negatively affects the overall accuracy. As such, the most common deployment approach of DNN-based applications on resource-constraint edge devices is the Remote-only Computing (RoC). With this, the network runs on the server side, and the input is directly transferred from the edge device to the server through a network connection. Then, the server computes the inferences and sends the output back to the device. However, such data transfer could lead to excessive latency times, especially in degraded channel conditions. As a compromise between the LoC and the RoC approaches, recently suggested *SC* frameworks propose to split DNN models into a head and a tail, deployed on edge device and server, respectively.

In this regard, our first contribution is [2], in which 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 further extended in [3], 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.

At the same time, current state-of-the-art approaches in different ML applications rely on advanced learning procedures, such as the Multi-Task Learning (MTL). In particular, MTL is a paradigm in which multiple related tasks are jointly learned to improve the generalizability of a model by using shared knowledge across different aspects of the input. This is achieved by jointly optimizing the model's parameters across all tasks, allowing the model to learn both task-specific and shared representations simultaneously. As a result, in [4], we propose, for the first time ever, how to partition multi-tasking DNN to be deployed within a SC framework, releasing the **MTL-Split** architecture. With this design, we can handle multiple tasks concurrently instead of the current focus on Single-Task Learning (STL) in SC, and through MTL, they increase task performance, overcoming the challenge of preserving only the performance of the main task.

III. Anomaly Detection

Surface Defect Detection (SDD) is a challenging problem in industrial scenarios, defined as the task of individuating samples containing a defect, i.e., samples that do not conform to a prototypical texture. In many real-world applications, a human expert inspects every product and removes those defective pieces. Unfortunately, humans are relatively slow in accomplishing this task, and their performances are subject to stress and fatigue.

Automated defect detection systems can easily overcome most of these issues by learning classifiers on defective and nominal training products. The main drawback is the data collection process required to train a model effectively. Indeed, defective items (i.e., positive samples) are relatively rare compared to nominal items (i.e., negative samples). Thus, the user may need to collect massive amounts of data to have enough positive samples.

To solve this issue, generative AI can represent a powerful tool for SDD, with defect image generation emerging as a promising approach to enhance detector performance. Thus, in [5], we propose a wild-and-creazy-idea to use a Diffusion Model for AD in Industry 4.0 processes. From this article, we have better understood the scientific challenges and formalized the problem rigorously.

Specifically, we can distinguish two different scenarios: *i)* when no defects are available (zero-shot data augmentation); *ii)* when some defects are available, which could be very few (few-shot, or N-shot with N small) or in a large number (full-shot or N-shot with N large). In the first case, a human-in-the-loop paradigm is employed. Specifically, a human operator can drive the generation of proper defects by exploiting their domain knowledge. This occurs using textual strings, which condition the generation of positive samples asking for specific defects (e.g., "scratches", "holes"). Instead, in the second scenario, when anomalous samples are available, fine-tuning can be done directly on them. In this case, human operators are unnecessary since the model can already learn what a defect looks like.

Due to the high complementarity of the two augmentation policies, we decided to use them together in a novel approach, dubbing **In&Out** [6] data augmentation, since it is a compromise between augmented images that are in and out-of-distribution. In particular, we show that In&Out-generated data allows the enrichment of the statistics of positive data (in-distribution), ameliorating the downstream classification performance in terms of recall.

This work was later extended in [6], in which we propose an interactive learning protocol where a vision language model is used to generate realistic images starting from textual prompts. Specifically, we promote using Denoising Diffusion Probabilistic Models (DDPMs) to produce fine-grained realistic defect images that can be used as positive samples to train an anomaly detection model. We name our approach **DIAG**, a training-free Diffusion-based In-distribution Anomaly Generation pipeline for data augmentation in the SDD task. By leveraging pre-trained DDPMs with multimodal conditioning, we can exploit domain experts' knowledge to generate plausible anomalies without needing real positive data. When using these augmented images to train an AD model, we show a notable increase in the detection performance compared to previous state-of-the-art augmentation pipelines. Furthermore, since we dive into spatial control approaches to enable the synthesis of defect samples, effectively utilizing domain expertise to generate more plausible in-distribution anomalies, we achieved high controllability and interpretability regarding the generated images.

IV. Future Developments

In the future, concerning SC, we aim to discuss their implications on controller design for Cyber-Physical System (CPS). Instead, regarding AD, we are planning further exploration across various datasets, particularly investigating how robust the image generation is compared to noisy textual prompts.

References

[1] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi, and M. Cristani, "A machine learning-oriented survey on tiny machine learning," *IEEE Access*, 2024.

[2] 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.

[3] 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.

[4] L. Capogrosso, E. Fraccaroli, S. Chakraborty, F. Fummi, and M. Cristani, "Mtl-split: Multi-task learning for edge devices using split computing," *arXiv preprint arXiv:2407.05982*, 2024.

[5] L. Capogrosso, A. Mascolini, F. Girella, G. Skenderi, S. Gaiardelli, N. Dall'Ora, F. Ponzio, E. Fraccaroli, S. Di Cataldo, S. Vinco *et al.*, "Neuro-symbolic empowered denoising diffusion probabilistic models for real-time anomaly detection in industry 4.0: Wild-and-crazy-idea paper," in *2023 Forum on Specification & Design Languages (FDL)*. IEEE, 2023.

[6] L. Capogrosso, F. Girella, F. Taioli, M. D. Chiara, M. Aqeel, F. Fummi, F. Setti, and M. Cristani, "Diffusion-based image generation for in-distribution data augmentation in surface defect detection," *arXiv preprint arXiv:2406.00501*, 2024.