9TH URV DOCTORAL WORKSHOP IN COMPUTER SCIENCE AND MATHEMATICS

Edited by
Alejandro Estrada, Oriol Farràs Ventura, Luis Pedro Montejano

UNIVERSITAT ROVIRA i VIRGILI

Departament d'Enginyeria Informàtica i Matemàtiques
[DΣIM]
Universitat Rovira i Virgili

Escola Tècnica Superior d'Enginyeria
Universitat Rovira i Virgili
ETSE

Title: 9th URV Doctoral Workshop in Computer Science and Mathematics
Editors: Alejandro Estrada, Oriol Farràs Ventura, Luis Pedro Montejano
April 2024

Universitat Rovira i Virgili
C/ de l'Escorxador, s/n
43003 – Tarragona
Catalunya (Spain)

ISBN: 978-84-1365-209-2
DOI: 10.17345/9788413652092

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

CONTENT

PREFACE

This book contains the abstracts of the works presented in the 9th Doctoral Workshop in Computer Science and Mathematics - DCSM 2024. It was celebrated in Universitat Rovira i Virgili (URV), Campus Sescelades, Tarragona, on April 23, 2024. The aim of this workshop is to promote the dissemination of ideas, methods, and results developed by the students of the PhD program in Computer Science and Mathematics from URV. It has been jointly organized by the research group of Discrete Mathematics and the Doctoral Program on Computer Science and Mathematics of Security of URV.

The editors and organizers invite you to contact the authors for more detailed explanations and we encourage you to send them your suggestions and comments that may certainly help them in the next steps of their PhD thesis. We thank all the participants and, especially, the students that presented their work in this DCSM workshop. Finally, we also want to thank Universitat Rovira i Virgili, the Departament d'Enginyeria Informàtica i Matemàtiques (DEIM), Escola Tècnica Superior d'Enginyeria (ETSE) and INCIBE for their support.

Alejandro Estrada, Oriol Farràs Ventura, Luis Pedro Montejano

Improving text anonymization via explainability of re-identification risk assessment

Benet Manzanares-Salor [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
benet.manzanares@urv.cat

Abstract. Text anonymization is usually approached as a named entity recognition (NER) task, but this often falls short in providing sound privacy protection. We introduce an approach designed to enhance the privacy safeguards provided by any text anonymization method –particularly those based on NER–, by leveraging the empirically observed re-identification risk and explainability methods. The proposed methodology allows users to intuitively set their preferred trade-off between privacy protection and utility preservation in terms of the probabilistic k-anonymity privacy model.

Keywords: Text anonymization; Neural language models; Explainability

1 Our proposal

We introduce a framework that enhances text anonymization by achieving a user-defined k-anonymity level. Starting with automatically anonymized documents (usually produced by NER-based methods), our approach enhances privacy by iteratively masking the terms contributing the most to re-identification. Our method is guided by a re-identification model [4]), which we use to assess the empirical re-identification risk of the documents to be protected, and an explainability method (such as SHAP [2]), which we use to detect the terms that contributed the most to that risk. The user defines the desired k value and the background knowledge that attackers can leverage to conduct re-identification attacks, and our framework iteratively masks documents' terms until meeting a probabilistic k-anonymity guarantee [7]. That is, when the re-identification probability of each protected document is below $1/k$.

[*] PhD advisor: David Sánchez

2 Results

Preliminary results show that our proposal significantly enhances the recall rate (that is, the privacy protection) of existing NER-based anonymization methods [3,5], closely aligning their privacy levels with those of state-of-the-art approaches [1,6] framed in the Privacy-Preserving Data Publishing (PPDP) field. Furthermore, our method operates automatically and without incurring in the substantial costs and complexities of these PPDP-oriented methods.

3 Future work

We plan to improve the runtime of our method by exploring alternatives to the explainability method we currently use (SHAP [2]), which is the most time-expensive component.

Acknowledgement. Partial support has been received from the Norwegian Research Council ("CLEANUP" project, nr. 308904), the INCIBE and European Union NextGenerationEU/PRTR (project "HERMES" and INCIBE-URV Cybersecurity Chair) and the University Rovira i Virgili and Banco Santander (2022PMF-BS-02).

References

[1] F. Hassan, D. Sánchez, J. Domingo-Ferrer. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering 35*, 1058–1071, 2023.

[2] S. M. Lundberg, S.-I. Lee. unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, Vol. 30, Association for Computing Machinery, pp. 4768–4777, 2017.

[3] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky. The stanford corenlp natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 55–60, 2014.

[4] B. Manzanares-Salor, D. S ánchez, P. Lison. Automatic evaluation of disclosure risks of text anonymization methods. *Privacy in Statistical Databases*, Vol. 13463 of Privacy in Statistical Databases, Springer, Paris, France, pp. 157–171, 2022.

[5] Microsoft Presidio `https://github.com/microsoft/presidio`

[6] A. Papadopoulou, P. Lison, L. Øvrelid, I. Pilán. Bootstrapping text anonymization models with distant supervision. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 4477–4487, 2022.

[7] J. Soria-Comas, J. Domingo-Ferrer. Probabilistic k-anonymity through microaggregation and data swapping. *2012 IEEE International Conference on Fuzzy Systems*, pp. 1–8, 2012.

Modeling Multicore Contention in Critical Real-Time Embedded Systems

Xavier Palomo Teruel [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
xavier.palomo@urv.cat

Abstract

In the context of real-time computing, the adoption of Commercially available Off The Shelf (COTS) multicore platforms is becoming prevalent, even within the most conservative domains. However, the transition to multicore systems has been met with the significant challenge of contention for shared hardware resources, which poses a risk to the predictability and performance essential in real-time applications. Addressing the complex interplay between multicore contention and system scheduling is crucial for these platforms to meet the rigorous demands of real-time computing.

This work investigates two approaches to compute the worst-case contention delay that could afflict a task due to the contention caused by simultaneous resource requests from co-running tasks on other cores. With a focus on statically scheduled systems, the research proposes two distinct models that operate under feasible assumptions to provide stringent upper bounds on contention delays. These models account for the intricacies of timing interference and the co-running task set dynamics, an aspect often overlooked in existing methodologies, and hence leading to more precise contention bounds than those provided by current state-of-the-art solutions.

1 Introduction

The advent of multicore systems in critical embedded real-time systems (CRTES) has necessitated a renewed focus on the timing analysis of systems operating on COTS multicore platforms. This thesis examines inter-core timing interference when concurrent applications access shared resources like buses on COTS multicore platforms, prevalent in domains demanding high safety and security. We propose analytical methods for establishing strict

[*] PhD advisor: Carlos Molina Clemente

worst-case delay bounds due to resource contention, with validation through a proof of concept on FPGA implementations in aerospace applications.

CRTES are integral to various domains, where failure can result in significant risks or losses. The development of CRTES must adhere to rigorous standards, such as DO-178B [1] and ISO26262 [2], with an emphasis on thorough verification and validation processes. These standards require solid timing guarantees, typically addressed through the derivation of Worst-Case Execution Time (WCET) bounds and subsequent schedulability analysis.

With CRTES facing an escalation in performance requirements and complexity, there is a shift from single-core to advanced multicore systems. The adoption of COTS multicore platforms is driven by their performance, cost-effectiveness, flexibility, and faster time to market, presenting a compelling alternative to custom hardware solutions [3].

2 System Model

The extent of contention a task faces is deeply rooted in the hardware and software setup of a system. Simplistic models often lead to unnecessary pessimism, prompting research that accounts for specific system and hardware conditions.

The interconnect, typically the main source of contention, is crucial when it allows parallel processing of multiple requests. This study focuses on systems where all external requests are managed through a single shared bus with a round-robin arbitration protocol, although our methods can be adapted for different system designs.

We examine a hardware architecture where cores possess private L1 caches and interface with main memory or a partitioned L2 cache via a common bus, minimizing indirect inter-core interference. Figure 1 illustrates this setup.

The caching strategy mirrors that of many embedded COTS platforms, with data caches using a write-through, no write-allocate policy, and L2 caches employing a write-back approach.

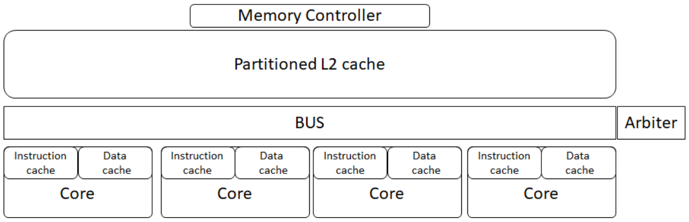

Fig. 1. Reference processor architecture

Our model accounts for the variability in request service times, which depend on the access type. The bus is the focal point of contention, with each request monopolizing it until serviced. Various access types, like L2 store (*s2h*) and load (*l2h*) hits, alongside clean (*l2mc*, *s2mc*) and dirty (*l2md*, *s2md*) misses, are orchestrated through this bus, with their occurrences logged by performance monitoring counters.

3 Proposed Methods

3.1 Integer Linear Programming

The ILP solution takes a strategic approach to the contention problem by utilizing the concept of *access pairing*. This approach models collisions between requests from different cores that contend for the same hardware resource, thereby affecting execution times. Pairing becomes crucial when tasks overlap in time across different cores, as depicted in Figure 2(a), where memory accesses from task τ_1 on Core 0 could collide with those from task τ_3 on Core 1. The delay induced by such contention is accounted for by pairing the accesses, with priority given to the most latency-inducing ones.

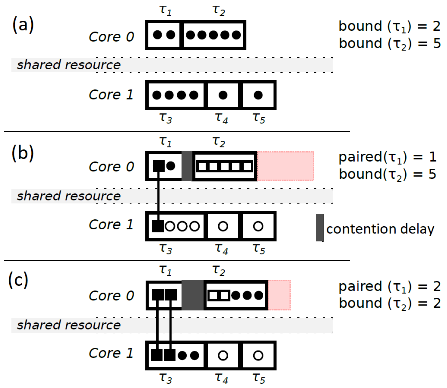

Fig. 2. Access pairing: motivation

This pairing mechanism links with task overlapping, creating a dynamic interplay where the worst-case pairing for an individual task may, paradoxically, lead to a reduced overall system makespan. As such, the ILP model facilitates a systematic computation of contention delays by carefully considering the mutual relationship between access pairing and task overlapping.

3.2 Iterative Approach

The iterative approach is implemented through a script that incrementally pairs accesses from co-running tasks, continuously refining the contention model. With each iteration, the script reassesses task alignments to update the computation of contention delays. This method's advantage lies in providing task-level guarantees, ensuring each task's timing requirements are individually met, albeit potentially at the expense of a larger overall makespan.

4 Partial Results

The graph presents a comparative analysis of execution times (ET) under different contention models. The observed execution time on the actual hardware (Board) is naturally consistently the lowest across all cores, indicating the real-world performance without any contention modeling.

The ILP-WCD approach, shown in green, and the Iterative approach, in yellow, both offer predictive models that consider various access types and task overlaps, resulting in higher ET than the Board. This increase accounts for the potential delays due to contention but maintains task-level guarantees.

The fTC (full Timing Composability) depicted in red, reflects the conservative estimate of WCET when assuming a uniform delay for all access types and neglecting the overlapping of tasks. This model yields the highest ET, as it overestimates the contention delays to ensure system safety.

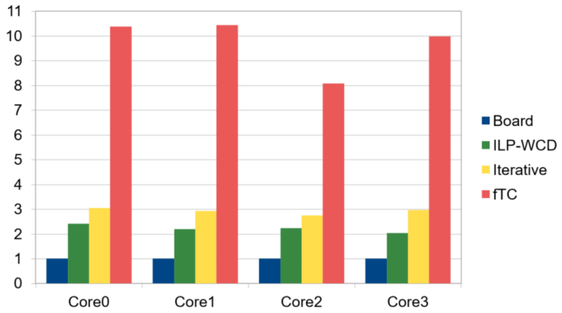

Fig. 3. Summary of the results

References

[1] RTCA and EUROCAE. DO-178C / ED-12C, Software Considerations in Airborne Systems and Equipment Certification. 2011.

[2] International Organization for Standardization. ISO/DIS 26262. Road Vehicles – Functional Safety. 2009.

[3] Intel Corporation. Intel® GO™ Automated Driving Solution Product Brief. `https://www.intel.es/content/dam/www/public/us/en/documents/platform-briefs/go-automated-accelerated-product-brief.pdf`.

Use of saliency maps in chemestry graph regression models

Natàlia Segura-Alabart [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
natalia.segura@urv.cat

1 Absrtact

This study presents a novel approach to enhance the interpretability of Graph Convolutional Networks (GCNs) in chemical compound analysis. Through the utilization of saliency mapping techniques, essential characteristics within molecular compounds, portrayed as graphs, are revealed. Experimental results demonstrate the model's predictive performance on chromatographic retention time datasets, showcasing its ability to accurately identify essential features for retention time prediction.

2 Introduction

A good prediction model has the ability to extract meaningful features in making predictions. However, the complexity of GCNs, characterized by numerous layers of non-linear transformations, poses challenges in interpreting model predictions. Consequently, saliency mapping techniques have been developed to explain the behaviour of such complex networks, including GCNs. These heatmaps serve to illustrate the relative importance of different segments within the input data, particulary graphs within the GCN framework, that contribute to the prediction of the particular property.

A molecular compound can be easily modeled as a graph consisting as atoms as the nodes and their bonds as the edges [1]. Saliency maps can be particular helpful in graphs, even more than for images, when identifying groups of atoms (a sub-graph structure on a molecular graph) that contribute to a particular property of a molecule.

This work aims to study regression models that uses GCNs and to evaluate the quality of saliency maps generated.

[*] PhD advisors: Francesc Serratosa and Alberto Fernández

3 Proposed method

In our method we have two steps, the creation of a regression model using GCNs and the evaluation of the saliency maps to asses its faithfulness. The model architecture is tailored for graph-based data processing. Input graphs, featuring node-specific attributes, are initially fed into the model. Subsequently, the data traverse through a series of five GCNs layers, each configured with a dimensionality of 256 and Rectified Linear Unit (ReLU) activation functions. Following the GCN layers, a readout operation, specifically a SegmentPoolingReadout, is performed to aggregate information across the graph structure. The aggregated features are then forwarded through two dense layers, each comprising 1024 units and employing ReLU activation functions. Finally, a single dense layer with one output unit is applied to produce the final prediction.

After the model training and testing phases, the final layer of the GCNs is utilized to generate saliency maps through the Grad-CAM method [2]. These maps enable the visualization of critical regions within the input graph that influence the model's predictions. To assess the interpretability of the model, we employ evaluation metrics to quantitatively measure the alignment between the highlighted graph segments and the actual molecular properties, thereby providing insights into the model's decision-making process and enhancing its transparency.

4 Experimental results and future work

Experiments were done using two chemical datasets, namely RPLC and HILIC [3], which correspond to distinct chromatographic separation modes: reversed-phase liquid chromatography and hydrophilic interaction liquid chromatography, respectively. The RPLC dataset comprised 852 graphs, while the HILIC dataset contained 1400 graphs. Both datasets quantify the retention time of each chemical compound using liquid chromatography techniques. The data from both datasets is split in 3 different datasets: training, validation and testing. We replicate the results 10 times for each dataset. Table 1 shows the predictive performance of the two datasets for both the training and testing sets in terms of Mean Square Error (MSE) and coefficient of determination (R^2).

Fig 1 shows two examples of chemical compounds represented as graphs with the gradient activation maps superimposed on the 2D structures. These two graphs came from the HILIC dataset. The green areas on the graphs represent certain features of the chemicals being studied, specifically the most polar parts. These polar parts are decisive for retention time prediction in HILIC techniques. The fact that the model correctly identifies these green areas shows that it is effectively learning and recognizing important features

Dataset		MSE	R^2
HILIC	Train	1.77	0.80
	Test	2.02	0.71
RPLC4	Train	0.60	0.84
	Test	0.66	0.81

Table 1: Predictive performance on the two chromatographic retention time datasets, HILIC and RPLC. MSE indicates mean square error and R^2 indicates the coefficient of determination. Both MSE and R^2 are the mean value of 10 replicates.

of the chemicals compounds. The next steps will focus on further evaluating the saliency maps faithfulness.

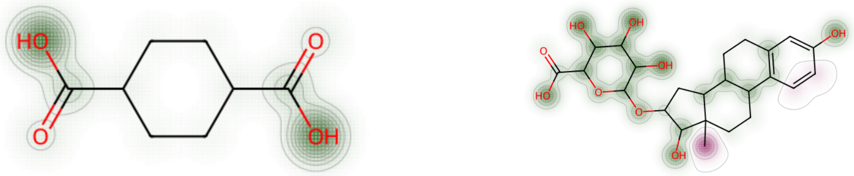

Fig. 1: Two examples of gradient activation maps generated from the HILIC dataset. The gradient activation maps (in green or purple) indicate how substructures contribute to retention (positively or negatively, respectively). The number of contour lines indicates to what degree the substructures contribute to retention, according to the model. The maximum number of contour lines is 10 per node.

Acknowledgement. This research is supported under the Martí i Franquès fellowship program(2020PMF-PIPF-45) and under the AGAUR research group 2021SGR-00111: ”ASCLEPIUS: Smart Technology for Smart Healthcare”, from Universitat Rovira i Virgili .

References

[1] Kim B-H and Ye JC Understanding Graph Isomorphism Network for rs-fMRI Functional Connectivity Analysis. *Front. Neurosci.*, 14:630,2020.

[2] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *ICCV 2017.*, 618-626,2017.

[3] A. Kensert, R. Bouwmeester, K. Efthymiadis, P. Van Broeck, G. Desmet and D. Cabooter. Graph Convolutional Networks for Improved Prediction and Interpretability of Chromatographic Retention Data. *Anal. Chem.*, 93(47):15633–15641,2021.

Exploring Graph Transformer Models for Molecular Regression Tasks

Sarah A. Fadlallah [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
sarah.fadlallah@urv.cat

Problem Statement

There are numerous methods of harnessing the rich features embedded within graphs. Graph Neural Networks (GNNs) have paved the way for these emerging variations. Graph Convolutional Networks (GCNs) [1], Graph Attention Networks (GATs), Gated Recurrent Networks (GRNs), and Graph Autoencoders (GAEs) have achieved state-of-the-art performance in graph prediction and generation tasks [2].

Among all these variants, identifying an appropriate embedding technique for a specific task can prove challenging. However, what is promising about this situation is that the adoption of various methods originally developed for different types of data has led to the discovery of powerful techniques. A vivid example of this is the utilization of Large Language Models (LLMs), which have expanded beyond their traditional role in natural language processing and sequence-to-sequence tasks to encompass other domains, such as image classification [4], audio processing [3], and graph embedding [5].

Objective

The goal is to investigate the potential contributions of attention models in graph learning and to pinpoint areas for further enhancement. This will entail experimenting with different graph transformer architectures and assessing the effectiveness of the resultant embeddings for a graph regression task using chemical datasets.

Methodology

Transformer architectures depend on the attention mechanism to compute the relationships between input components. This yields an attention score that

[*] PhD advisor: Francesc Serratosa, Carme Julià

can be utilized to predict the next token in sequence-to-sequence problems. The attention between two tokens in an input sequence can be calculated as follows 1,

$$A(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

In the first selected graph transformer model [6], the node feature is encoded with its in and out degrees.

$$h_i^{(0)} = x_i + z^-_{degz^-(v_i)} + z^+_{degz^+(v_i)} \tag{2}$$

The self-attention mechanism has been modified to incorporate additional graph properties denoted by bias terms that encode the shortest path distance between nodes and the averaged dot-product path as additional edge features.

$$A_{ij} = \frac{(h_i WQ)\,(h_j WK)\,T}{\sqrt{d}} + b_{\phi(v_i,v_i)} \tag{3}$$

On the other hand, the architecture presented in [7] was formed by injecting the laplacian eigenvectors of a graph as positional encoding, adding it to the node representation as per the equation.

$$\lambda_i^0 = C^0\lambda_i + c^0; h_0^i = \hat{h_0^i} + \lambda_0^i \tag{4}$$

Where the node features are linearly projected through a learned embedding.

Preliminary results and future directions

Different ablation studies were conducted to determine the effect of the aforementioned added attributes. Predicting a global property of a given graph had a noticeable increase in accuracy when using degree-augmented node representation along with path and edge information in the Graphormer architecture and using the Laplacian eigenvectors in the Graph transformer one. Embeddings from both methods have yielded promising results in predicting the solubility in the ESOL dataset as shown in figure 2.

The next steps will focus on improving the prediction accuracy, and testing the resulting embeddings for graph generation tasks.

Acknowledgement. This research is supported by the Martí Franquès program in addition to AGAUR research group 2021SGR-00111: ”ASCLEPIUS: Smart Technology for Smart Healthcare”.

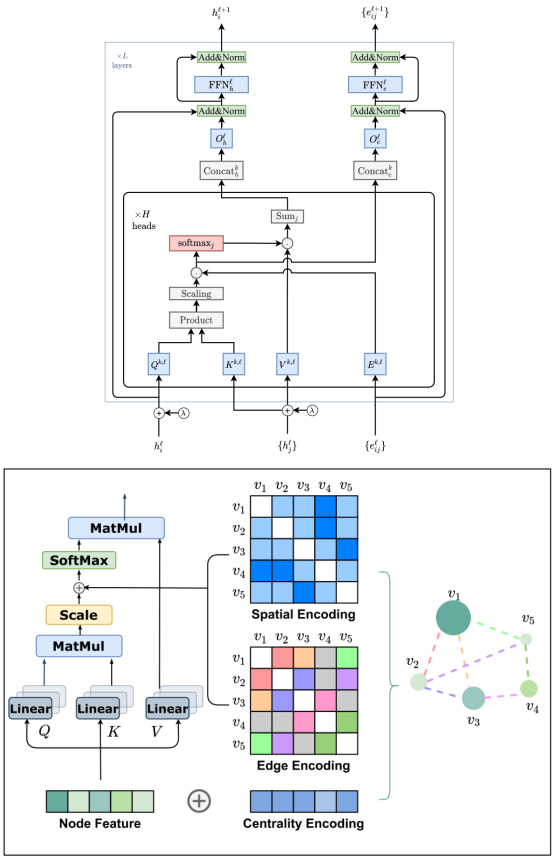

Fig. 1. An illustration of the two chosen architectures for analysis Graph Transformer (left), Graphormer (right).

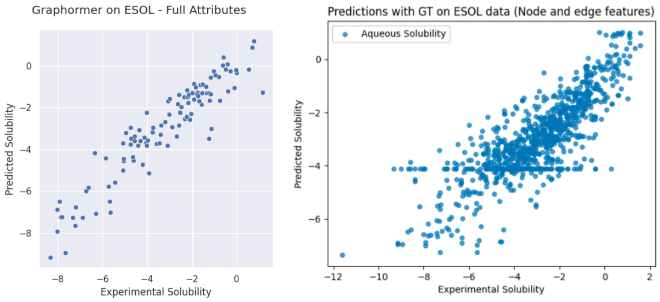

Fig. 2. Scatter plots showing both the GT and Graphormer to predict solubility on the ESOL dataset.

References

[1] Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR*, abs/1609.02907, 2016.

[2] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. & Sun, M. Graph neural networks: A review of methods and applications. *AI Open*. **1** pp. 57-81 (2020), https://www.sciencedirect.com/science/article/pii/S2666651021000012

[3] Liu, X., Lu, H., Yuan, J. & Li, X. CAT: Causal Audio Transformer for Audio Classification. (2023)

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*. **abs/2010.11929** (2020), https://arxiv.org/abs/2010.11929

[5] Yun, S., Jeong, M., Yoo, S., Lee, S., Yi, S., Kim, R., Kang, J. & Kim, H. Graph Transformer Networks: Learning Meta-path Graphs to Improve GNNs. *CoRR*. **abs/2106.06218** (2021), https://arxiv.org/abs/2106.06218

[6] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y. & Liu, T. Do Transformers Really Perform Bad for Graph Representation?. *CoRR*. **abs/2106.05234** (2021), https://arxiv.org/abs/2106.05234

[7] Dwivedi, V. & Bresson, X. A Generalization of Transformer Networks to Graphs. *CoRR*. **abs/2012.09699** (2020), https://arxiv.org/abs/2012.09699

[8] Le, T., Le, N., and Le, B. Knowledge graph embedding by relational rotation and complex convolution for link prediction. *Expert Systems with Applications*, 214:119122, March 15 2023. doi: 10.1016/j.eswa.2022.119122. ISSN: 0957-4174, EISSN: 1873-6793.

Cost-efficient Stream Processing with Serverless Functions

Pablo Gimeno Sarroca *

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
pablo.gimeno@urv.cat

1 Introduction

Despite that Function-as-a-Service (FaaS) has settled down as one of the fundamental cloud programming models, it is still evolving quickly. Recently, Amazon has introduced S3 Object Lambda, which allows a user-defined function to be automatically invoked to process an object as it is being downloaded from S3. As with any new feature, careful study thereof is the key to elucidate if S3 Object Lambda, or more generally, if inline serverless data processing, is a valuable addition to the cloud. For this reason, we conduct an extensive measurement study of this novel service. We particularly focus on the streaming capabilities of this new form of function, as it may open the door to empower existing serverless systems with stream processing capacities. We discuss the pros and cons of this new capability through several workloads, concluding that S3 Object Lambda can go far beyond its original purpose and be leveraged as a building block for more complex abstractions.

2 Background and motivation

Function-as-a-Service (FaaS) is a popular cloud computing model, where applications leverage cloud resources via a series of user-defined functions that encapsulate the application code. By controlling all aspects of function execution, cloud platforms deliver a serverless computing model where users do not have to explicitly provision and administer cloud resources. Moreover, FaaS employs a fine-grained pricing model allowing applications to pay for the resources they utilize, while giving the illusion of near-infinite horizontal scaling. FaaS platforms are available in all major cloud providers, and have been pulled for a variety of applications such as web services, parallel and distributed computing [4], and even machine learning (ML) [3].

* PhD advisor: Marc Sánchez-Artigas

Recently, Amazon introduced S3 Object Lambda [1], a new capability to better support one of the most common serverless patterns as of today: the processing of data retrieved from the S3 storage service before handing it to an application. In other words, S3 Object Lambda allows developers to pipe a file through a Lambda invocation on read (i.e., as part of the lifecycle of S3 `GetObject` request), so that applications will read that file through the "lens" the user crafted for it. Although this is a very exciting new capability, it does not mean that developers will need it. For this reason, it is felt useful not only to characterize the performance of this new service, but to determine for what uses cases it is ideal and which paths forward can be envisaged.

In this work we present the highlights of our thorough study of S3 Object Lambda [5], which provides valuable information of its architecture and performance, proves the new service's utility in a variety of use cases and enables our current line of research.

3 Study results

The study is divided in three parts. In the first part, we elucidate some basic details about the architecture and resource management of S3 Object Lambda. Secondly, we characterize its performance obtaining several relevant metrics. Finally, we assess the streaming capabilities of S3 Object Lambda by running multiple workloads. The following subsections contain a brief summary of the results.

3.1 Architecture

A1: **S3 Object Lambda is not a near-data-processing accelerator**: contrary to what one would expect from such a service, S3 Object Lambda functions do not run on a dedicated cluster closer to storage, but rather in the same cluster as regular Lambda functions.

A2: **Inconsistent enforcement of S3 Object Lambda quotas**: S3 Object Lambda functions have a maximum 60-second timeout. This, however, only applies to the S3 Object Lambda endpoint, which means that a function can only stream data back to the client during the first 60 seconds of execution. However, a user could implement a function that only emits data during the first 60 seconds but inadvertently runs for longer, as a regular Lambda's timeout is 15 minutes, incurring unexpected costs.

A3: **Role of the S3 Object Lambda Access Point**: the additional components of S3 Object Lambda have an impact in the function's latency (startup time and time to first byte). The idea of gluing both S3 and Lambda is neat, but introduces performance penalties and quota inconsistencies.

3.2 Performance measurements

- **P1: Overhead analysis**: S3 Object Lambda's TTFB times are always longer than those of Lambda, and up to 1.8x higher. This is due to the extra S3 Object Lambda Access Point, which introduces an extra hop between the client and the function.
- **P2: Cold start**: As in [6], a cold start refers to the process of launching a new function instance. We observe that S3 Object Lambda's cold start time is not much longer than Lambda's, up to 1.2x, which is great news considering the aforementioned extra component.

3.3 Serverless streaming use cases

- **S1: Memory-optimized streaming**: since Object Lambda enables inline processing where I/O and compute operations overlap, it allows masking access latency to S3, while optimizing function memory resources, as data can be processed in small batches. We implemented two memory-optimized use cases: `Grep` and `Zlib` decompression, both in Lambda and Object Lambda. The results of these jobs highlight the importance of choosing the correct use cases for Object Lambda, since use cases such as zip decompression, where there is no data reduction in the Object Lambda functions, can lead to high data transfer costs. However, use cases like `Grep` prove that S3 Object Lambda can process the same amount of data as Lambda with similar execution times, but smaller memory configurations, thus significantly reducing costs.
- **S2: Streaming pipelines**: Object Lambda functions can be chained together creating a pipeline of streaming operations. We chose to perform a Parallel Tree Reduction as an example of this feature. We compared a streaming, Object-Lambda-based implementation with a batch, Pywren-based [4] implementation of the pipeline. The results show that the streaming implementation is not only faster (1.27x), but also slightly cheaper (1.15x), as the streaming implementation is capable of pipelining I/O and computation and performs a significant data reduction in the Object Lambda layers, thus reducing the cost of passing intermediate data between chained Object Lambdas. Additionally, the streaming implementation is capable of producing preliminary results as the job progresses, producing almost perfectly accurate results 3.75x faster than the batch implementation.

4 Future work

The results obtained in our study prove the utility of inline processing in FaaS and open up the possibility of leveraging S3 Object Lambda not only as a mere tool for transforming outgoing data from S3, but also as the building block for more complex streaming use cases. As such, we are currently exploring the integration of streaming-capable FaaS with state-of-the-art stream processing engines (SPE), such as Apache Flink [2]. Our current research line aims at improving cost-efficiency of stream processing pipelines by aiding SPEs with a layer of serverless compute resources.

5 Conclusion

In this work, we have presented the highlights of the first in-depth study of S3 Object Lambda [5]. Specifically, we have provided insights into its architecture, resource utilization and performance. But most importantly, we have surfaced the novel opportunities brought about by this brand new inline data processing service through several workloads. We conclude that S3 Object Lambda can go far beyond its original aim and be leveraged as a building block for more complex streaming abstractions.

Acknowledgement. This research work has been partially supported by the EU through the projects Extract (No. 101092749), NearData (No. 101092644) and CloudSkin (No. 101092646), and the Spanish Government under grant agreement PID2019-106774RB-C22. Pablo Gimeno Sarroca is a Martí Franquès Research Grant Fellow — Banco Santander Ed.

References

[1] AWS News Blog. Introducing amazon s3 object lambda – use your code to process data as it is being retrieved from s3. `https://aws.amazon.com/blogs/aws/introducing-amazon-s3-object-lambda-use-your-code-to-process-data-as-it-is-being-retrieved-from-s3/`, Mar. 2021.

[2] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. Apache flink: Stream and batch processing in a single engine. *The Bulletin of the Technical Committee on Data Engineering*, 38(4), 2015.

[3] Pablo Gimeno Sarroca and Marc Sánchez-Artigas. Mlless: Achieving cost efficiency in serverless machine learning training. *Journal of Parallel and Distributed Computing*, 183:104764, 2024.

[4] Eric Jonas, Qifan Pu, Shivaram Venkataraman, Ion Stoica, and Benjamin Recht. Occupy the cloud: distributed computing for the 99%. In *2017 ACM Symposium on Cloud Computing (SoCC'17)*, pages 445–451. ACM, 2017.

[5] Pablo Gimeno Sarroca and Marc Sánchez-Artigas. On data processing through the lenses of s3 object lambda. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pages 1–10, 2023.

[6] Liang Wang, Mengyuan Li, Yinqian Zhang, Thomas Ristenpart, and Michael Swift. Peeking behind the curtains of serverless platforms. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, pages 133–146, 2018.

Challenges of Serverless Scientific Data Analysis: Genomics Variant Calling Use-Case

Aitor Arjona [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
aitor.arjona@urv.cat

1 Introduction

With the escalating complexity and volume of genomic data, the capacity of biology institutions' HPC faces limitations. While the Cloud presents a viable solution for short-term compute and storage resource elasticity, its intricacies pose challenges for bioinformatics users. Alternatively, serverless computing allows for workload scalability with minimal developer burden. However, porting a scientific application to serverless is not a straightforward process. This work presents a Variant Calling genomics pipeline migrated from single-node HPC to a serverless architecture. We outline the inherent challenges of this approach and the engineering efforts required to achieve scalability with current state-of-the-art serverless platforms. The insights obtained in this work determine the future challenges to be addressed for making scientific applications in serverless environments more efficient and cost-effective. We contribute by open-sourcing the pipeline for future systems research and as a scalable user-friendly tool for the bioinformatics community.

2 Context and motivation

Genome sequence analysis is a compute- and data-intensive task. Current sequencing techniques generate an overwhelming volume of data, with repositories accumulating several petabytes each year. The European Nucleotide Archive alone surpassed 40 petabytes of data in 2021, and this trend is expected to continue in the future [1]. Consequently, institutions struggle to meet the ever-growing demands for genomics workloads. Although High-Performance Computing (HPC) installations are commonplace in biology departments, they may fall short in providing sufficient capacity to handle peak demands, reanalyze extensive public datasets generated by large consortia, or minimize the run time needed for specific analyses.

[*] PhD advisors: Pedro García-López and Marc Sanchez-Artigas

3 Problem statement

Faced with this situation, many institutions often turn to public Clouds in seek of compute and storage resource elasticity. The Cloud allows to meet exceptional capacity needs on demand while avoiding the burden of acquiring and maintaining physical infrastructure on a permanent basis. Although there are genomics-centered distributed computing frameworks available for the Cloud (e.g., Nextflow [3] or GATK for Apache Spark), they present numerous challenges for bio-informatics users. Evaluating, configuring, deploying and scaling the necessary resources and services for each application becomes a daunting task due to the multitude of heterogeneous Cloud providers and the complexities of their services.

Function-as-a-Service (FaaS) offers a compelling alternative for highly parallel data-intensive applications [5,6,7], with minimal configuration burden and the ability to instantly scale up and *down to zero*, making it an ideal choice for less experienced Cloud users to deploy workloads. Recent advancements in FaaS services, such as Amazon Lambda's increased resource allocation[2], open up new possibilities for tackling demanding workloads with serverless. However, adopting the serverless paradigm, despite its cost-effectiveness and scalability benefits, introduces its own set of challenges [4]. While some efforts have been made to ease the transition from existing single-machine code to distributed and serverless code [8], it remains a complex task, often necessitating substantial application re-architecting.

4 Case study: Genomics Variant Calling

Here we present a case study involving the porting of a genomics Variant Calling workflow from HPC to a serverless architecture. Our focus lies on studying the key challenges that arise from this approach, particularly those concerning concurrency and parallelism, input data partitioning, and stateful data movements. The objective of our work is to harness the massive and instant parallelism of current state-of-the-art FaaS services [2] to achieve high scalability in a genomic variant calling pipeline, all while ensuring a streamlined experience with minimal burden for the end user. Despite the large data movements involved, by decomposing the workflow into fine-grained tasks and leveraging parallelism in multi-CPU functions, we can ultimately achieve high performance and scalability in a cost-effective way.

Our primary contribution lies in the insights gained from the engineering efforts undertaken during the migration. The key challenges outlined below are proposed as targets for future research, with the aim of providing novel

[2] `https://aws.amazon.com/about-aws/whats-new/2020/12/aws-lambda-supports-10gb-memory-6-vcpu-cores-lambda-functions/`

solutions that contribute to improving the performance of serverless scientific applications. In particular, those challenges are:

1. **Data partitioning:** Unlike cluster-oriented frameworks, serverless provides fine-grained execution of tasks, which allow to substantially increase the parallelism. This makes data partitioning crucial to leverage the parallelism of FaaS. However, *many scientific data formats are not inherently prepared to be read in dynamically-sized logical chunks. Consequently, large files often need to be split into multiple smaller ones, creating a bottleneck in serverless workflows.*
2. **Orchestration:** FaaS platforms limit the number of functions running concurrently, meaning that functions are note preemtive. Users cannot initiate new function executions once the limit is reached. Consequently, *processes that are blocking in local multi-threaded non-distributed applications are incompatible with serverless architectures. As a result, all tasks must be asynchronous, requiring an orchestration component to execute functions at the appropriate times.*
3. **Stateful data movements:** Serverless functions are stateless and cannot open communication channels between them. This requires an external storage layer to pass data between functions for stateful data movements (e.g., data shuffle). Following the state-of-practice, object storage provides scalability and low per-GB storage costs, but due to its high latency, I/O time billed at serverless function runtime results in overall high costs. *A high-performance middleware for indirect communication is necessary to provide stateful data movements for stateless serverless functions, with the aim of minimizing the latency to reduce function runtime.*

5 Conclusion

In this work, we present a genomics variant calling pipeline ported from single-node HPC to serverless Cloud with the aim of scalability. The serverless architecture allowed us to scale a realistic benchmark of variant calling in the field of human genomics, with more than 1200 parallel tasks, achieving an execution time of 7 minutes and 16.57 seconds. For reference, the same workload ran for ≈ 14 hours on the HPC setup. The total execution cost was 10.68 USD for AWS Lambda GB-sec billing.

In summary, serverless enables to massively scale scientific workloads, but adapting them from HPC requires thorough adjustment and complex workflow restructuring. Porting this pipeline to a serverless architecture has allowed us to determine the key challenges that must be addressed in order to improve the performance of these applications in the Cloud, prompting to future research directions. We've open-sourced this work, on GitHub[3], both

[3] https://github.com/CLOUDLAB-URV/serverless-genomics-variant-calling

for the evaluation of serverless research systems and as an open-source tool for scalable bioinformatics variant calling in the Cloud.

Acknowledgement. This work has been partially funded by the European Union through the Horizon Europe NEARDATA (101092644), CloudSkin (101092646), CLOUDSTARS(101086248), EXTRACT (101093110) projects; Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU, (within the framework of the PRTR and the MRR) through the CLOUDLESS UNICO I+D CLOUD 2022. Aitor Arjona is a URV Martí Franquès grant fellow.

References

[1] European Nucleotide Archive. Statistics. `https://www.ebi.ac.uk/ena/browser/about/statistics`, 2022.

[2] Daniel Barcelona-Pons and Pedro García-López. Benchmarking parallelism in faas platforms. *Future Generation Computer Systems*, 124:268–284, 2021.

[3] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

[4] Joseph M. Hellerstein, Jose Faleiro, Joseph E. Gonzalez, Johann Schleier-Smith, Vikram Sreekanti, Alexey Tumanov, and Chenggang Wu. Serverless computing: One step forward, two steps back, 2018.

[5] Eric Jonas, Qifan Pu, Shivaram Venkataraman, Ion Stoica, and Benjamin Recht. Occupy the cloud: Distributed computing for the 99%. In *Proceedings of the 2017 Symposium on Cloud Computing*, SoCC '17, page 445–451, New York, NY, USA, 2017. Association for Computing Machinery.

[6] Qifan Pu, Shivaram Venkataraman, and Ion Stoica. Shuffling, fast and slow: Scalable analytics on serverless infrastructure. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, pages 193–206, Boston, MA, February 2019. USENIX Association.

[7] Vaishaal Shankar, Karl Krauth, Qifan Pu, Eric Jonas, Shivaram Venkataraman, Ion Stoica, Benjamin Recht, and Jonathan Ragan-Kelley. numpywren: serverless linear algebra, 2018.

[8] Josef Spillner, Cristian Mateos, and David A. Monge. Faaster, better, cheaper: The prospect of serverless scientific computing and hpc. In Esteban Mocskos and Sergio Nesmachnow, editors, *High Performance Computing*, pages 154–168, Cham, 2018. Springer International Publishing.

Automated region of interest localization on breast histological images

Alessio Fiorin *

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
Pathology Department, Oncological Pathology and Bioinformatics Research Group,
Hospital de Tortosa Verge de la Cinta, ICS, Institut d'Investigació Sanitària Pere
Virgili (IISPV), 43500 Tortosa, Tarragona, Spain
alessio.fiorin@estudiants.urv.cat

1 Introduction

Histological assessment is vital in diagnosing, prognostic, and treating diseases. With the continuous progress of scientific research, computational pathology has always a more crucial role in revolutionising conventional clinical workflows within pathology laboratories [1]. Thanks to computer vision and digital pathology advancements, pathologists can now analyse digitised glass slides, commonly known as whole slide images (WSIs), with the help of computer-aided pathology tools [2].

A way to mitigate bias and enhance the robustness of predictive biomarker assessment is to assess multiple WSIs from the same specimen. The aforementioned multiple assessment is thus helpful, for instance, to enhance the robustness of the evaluation of tumour-infiltrating lymphocytes (TILs), which are crucial to understanding tumour characteristics pertinent to prognosis and treatment planning in breast cancer [3].

The main aim of this study is to optimise clinical workflow within pathology laboratories for TIL assessment, automating the identification of ROIs for assessing TIL scoring variability across consecutive specimen cuts. The ROIs are derived from the ROI of a cut selected by a pathologist following Salgado's criteria [4] using a WSI registration-based technique resilient to variations inherent in the tissue preparation process.

The study will then compare the ROIs computed using the registration technique and those annotated by the pathologist, complemented by qualitative evaluations through blind assessments conducted by another pathologist.

The registration of WSIs is a challenging task since it deals with gigapixel scale images, inherent variations in appearance, structure, and morphology

* PhD advisors: Prof. Carlos López Pablo, prof. Marylène Lejeune, prof. Hatem Rashwan, and prof. Domenec Puig

among tissue regions in non-consecutive sections, and the presence of artefacts, tears, and deformations due to the preparation of the tissue blocks [1].

2 Data collection and preparation

We analyze eleven breast cancer specimen cases, each comprising three consecutive cuts of 2 µm of thickness and stained with hematoxylin and eosin, successively digitized using a PANNORAMIC 250 Flash 3DHistech scanner. A pathologist, following the Salgado's criteria [4], searched the more adequated area for the evaluation of TILs. After finding the best area, the expert annotates the same ROI in the three consecutive cuts of each specimen using SlideViewer, a proprietary software developed by 3DHistech.

The annotation had a rectangle shape with size 4182 × 7892 and subsequently, we convert these annotations into GeoJSON format, which is compatible with QuPath, an open-source platform used in digital pathology. This study has the approval of the Ethics Committee of the Pere Virgili Research Institute (reference number 128/2022).

3 Methodology

A pair of WSIs was aligned in two sequential steps: a preprocessing phase followed by rigid registration. To address the challenge of managing gigapixel-sized images, we employed a downscaled version of the WSIs, reducing their shape by 16. The preprocessing phase targets specific areas via initial tissue segmentation, enabling focused registration only on the detected tissue regions. For this purpose, we employed the FESI method, a computationally efficient approach committed to histological WSIs [5].

After preprocessing the downscaled WSIs, we applied rigid registration, aligning the image with the image of the first cut through rotation and translation, thus implementing an Euclidean transformation. The rigid registration ensures that when computing the vertices of the ROI, we obtained a rectangle with an identical shape to the ROI in the reference image. The library used for implementing the rigid registration is called ANTsPy.

After computing the ROI on the downscaled WSI, we calculated the ROI on the original image by multiplying the coordinates on the downscaled WSI by a factor of 16. The registration is robust since it considers variations in the reference image, possibly being provided in a flipped orientation due to the tissue preparation phase.

Two registration approaches were compared: the first one makes a direct prediction of the ROIs of the second and third cuts from the first cut and the second one makes a chain approach where the ROI of the third cut is derived from the predicted ROI of the second cut. Thus the ROI's prediction in the second cut is the same for both methodologies, while change in the ROI's prediction in the third cut.

Quantitative agreement is assessed using Intersection over Union (IoU), used to measure the similarity between the predicted ROI and the one annotated by the pathologist.

Regarding qualitative assessment, an independent pathologist decides blindly between the ROI annotated by the pathologist and the one computed by the registration algorithm.

The motivation behind the development of these two approaches is to establish whether aligning the third cut with the first cut is sufficient for achieving accurate alignment or if a chain approach utilizing a reference image closer in physical distance to the cut to be aligned is required. Notably, employing the first approach ensures the correctness of the ROI in the reference image, as any registration errors in predicting the second cut's ROI could propagate if the prediction fails.

4 Results

Regarding quantitative assessment, IoU scores for ROI predictions of the second cut have an average of 0.823 +/- 0.063, which indicates a strong alignment between machine predictions and pathologist annotations. When considering ROI predictions for the third cut, the first approach shows an IoU score with pathologist annotations at 0.772 +/- 0.073, while the second approach exhibits a score at 0.776 +/- 0.063.

On a qualitative level, pathologist preferences exhibit variability. Concerning ROI predictions for the second cuts, the pathologist favours five machine predictions, four pathologist annotations and two no discernible differences between the machine and pathologist. Regarding ROI predictions for the third cuts, the first approach receives preference from pathologist annotations in almost all the cases (9/11), with the remaining two indicating an equal preference. After that, the second approach shows three favouring machine predictions, five cases preferring pathologist annotations, and the other three expressing no distinct preference.

5 Discussion

The study shows the usefulness of a robust, rigid registration technique for locating the ROI across specimen cuts. The chain methodology shows enhanced agreement with pathologist annotations, indicating its superiority in accurately predicting the ROI for the third cuts. This advantage derives from the intrinsic similarity in structure and morphology among tissue regions in consecutive sections.

However, the first approach retains its significance as a viable fallback option. It serves as a trustworthy backup in scenarios where the prediction of the ROI in second cuts encounters difficulties, effectively preventing the propagation of errors to subsequent forecasts in this study for the third specimen cut. No flipped images were detected, which means that there were no errors during the tissue preparation phase.

6 Conclusions

The study underscores the importance of accurate ROI prediction for TIL scoring consistency. The robust, rigid registration technique offers a promising approach for this purpose, with implications for improving pathology workflows and enhancing diagnostic accuracy in breast cancer assessment.

Acknowledgement. This work was supported by BosomShield, a project that has received funding from Marie Skłodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01) under grant agreement 101073222. Further support was provided by SCARLET, a project funded by Proyectos Estratégicos Orientados a la Transición Ecológica y a la Transición Digital, from the 2021 call of the Ministerio de Ciencia e Innovación, with grant number TED2021-130081B-C22 and funding from NextGenerationEU.

Special thanks to the pathology department at the Hospital of Tortosa Verge de la Cinta (HTVC) for helping with data collection and the qualitative assessment phases.

References

[1] Weitz, P., Valkonen, M., Solorzano, L. et al. A Multi-Stain Breast Cancer Histological Whole-Slide-Image Data Set from Routine Diagnostics. *Scientific Data.* **10** (2023,8), http://dx.doi.org/10.1038/s41597-023-02422-6

[2] Nam, S., Chong, Y., Jung, C., et al. Introduction to digital pathology and computer-aided pathology. *Journal Of Pathology And Translational Medicine.* **54**, 125-134 (2020)

[3] Adams, S., Gray, R., Demaria, S., et al. Others Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. *Journal Of Clinical Oncology.* **32**, 2959 (2014)

[4] Salgado, R., Denkert, C., Demaria, S. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Annals Of Oncology.* **26**, 259-271 (2015)

[5] Bug, D., Feuerhake, F., Merhof, D. Foreground extraction for histopathological whole slide imaging. *Bildverarbeitung Für Die Medizin 2015: Algorithmen-Systeme-Anwendungen. Proceedings Des Workshops Vom 15. Bis 17. März 2015 In Lübeck.* pp. 419-424 (2015)

This book contains the proceedings of the 9th Doctoral Workshop in Computer Science and Mathematics - DCSM 2024. It was celebrated in Universitat Rovira i Virgili (URV), Campus Sescelades, Tarragona, on April 23, 2024. The aim of this workshop is to promote the dissemination of ideas, methods, and results developed by the students of the PhD program in Computer Science and Mathematics from URV.

Departament d'Enginyeria [DΣIM] Informàtica i Matemàtiques
Universitat Rovira i Virgili

Escola Tècnica Superior d'Enginyeria
Universitat Rovira i Virgili
ETSE