Machine Learning in Medical Imaging

Martin Bobák[1], Ladislav Hluchý[1], Mara Graziani[2], Henning Müller[2]

[1] Institute of Informatics, Slovak Academy of Sciences
Bratislava, Slovakia
{martin.bobak, ladislav.hluchy}@savba.sk

[2] University of Applied Sciences Western Switzerland (HES-SO)
Sierre, Switzerland
{mara.graziani,henning.mueller}@hevs.ch

Abstract
The paper describes the machine learning in medical imaging which represent one of the exascale services prepared in the PROCESS project. It also presents the architecture capable to handle such services with quantitative analysis performed at two computing sites.

Keywords **– machine learning, deep learning, exascale architecture, medical imaging, high-performance computing, cloud computing.**

1. Introduction

Digital histopathology is the automatic analysis of a biopsy or surgical tissue specimens that are captured by a high-resolution scanner and stored as Whole Slide Images (WSIs). WSIs are usually stored in a multi-resolution pyramid structure, where each layer contains down-sampled versions of the original image. The amount of information in WSIs is large, since it includes tissue that is not relevant for cancer diagnosis (e.g. background, healthy tissue, etc.) For this reason, machine learning and deep learning models are built to detect Regions of Interest (ROIs) within WSIs. ROIs are portions in the WSI where the cancer is visible and therefore contain relevant information to train the network.

2. Use case description

Figure 1 shows a data pre-processing pipeline. As a first step the raw WSIs are analysed at a very low resolution, and tissue is filtered from the background. Based on physician's annotations, tumor regions are isolated. These regions represent ROIs that are used to perform network training. From the normal tissue and from tumor ROIs, patches are extracted at a higher level of magnification. Higher resolution highlights qualitative features of the nuclei which are essential for cancer detection. For instance, recent research has shown that performance of classifiers improves with higher resolution patches.

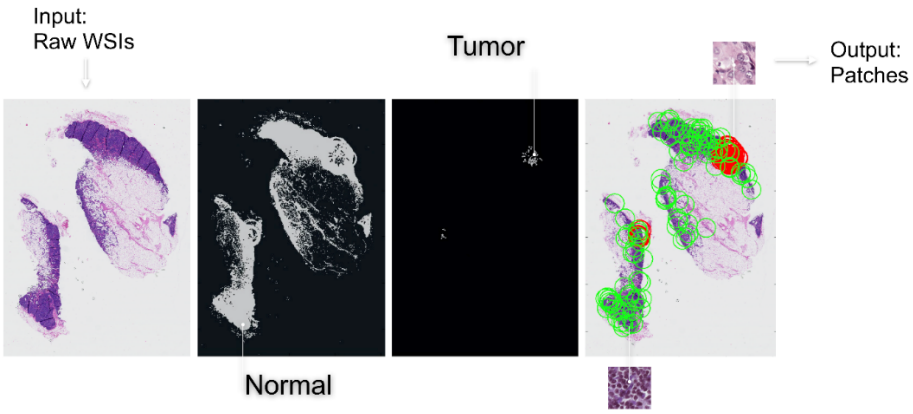

Figure 1: WSIs preprocessing pipeline - Normal tissue and tumor tissue masks are extracted and high-resolution patches are sampled from the selected regions.

The use case application is organized in three layers. Layer I implements the extraction of patches of dimensions 224x224 pixels from the gigapixel slides of breast lymph node

tissue. Patches are randomly sampled from the slide, in which areas of tumour were annotated by a physician. Patches belonging to a tumorous region are assigned a 'tumor' label (a Boolean variable equals to true). The extracted data are stored in an intermediate dataset with the corresponding labels. Layer II loads the intermediate dataset of patches and labels and trains a state-of-the-art deep convolutional network to classify the two patch types. Different models can be chosen by a configuration parameter. Layer III focuses on network robustness and interpretability. A summary of the use case application layers can be found in Table 1.

Table 1: Camnet - Interchangeable networks for Camelyon.

Layer I: Data pre-processing and patch extraction	**Layer II: Local and distributed training**	**Layer III: Performance boosting and interpretability**
Creation of normal and tumour tissue masks from the physician's annotations	State of the art deep convolutional networks currently implemented: Resnet50, Resnet101, InceptionV3	Generation of intermediate visualizations
Random sampling of high-resolution patches and labelling	Local training on single and multiple GPUs	Feature importance analysis
Intermediate storage of the patches on H5DS	Training on HPC clusters	Perturbation robustness analysis
	Distributed training	

3. PROCESS platform

The reference exascale architecture (see Figure 2) is divided into the following parts (from top to bottom):

- **Users of the exascale scientific applications** (in yellow) - the exascale system has to support functionalities required by its user communities. The best way is to build it on containerization. All of the applications are stored in a containerized repository that is available to use communities.
- **Virtualization layer** (in blue) - is situated between the containerized application repository and platform infrastructure managers. Interoperability of data and computing infrastructures is the key and critical requirement of the exascale systems.
- **Data management** (in green) - could be divided into two main groups: distributed data federation and metadata. The metadata module has to be federated and distributed as well as the management system for the data infrastructure itself. At this level of the infrastructure, the system architect has to be careful whether the component will be containerized, or not. Micro-services serve as adapters and connectors to infrastructural services. They are integrated into a containerized micro-infrastructure, which is customized according to requirements coming from a use case and connecting them to a distributed virtual file system.
- **Computing Management** (in red) - this part of the infrastructure is related to scheduling and monitoring computing resources. Two kinds of resources are supported, namely: high-performance computing (HPC) resources, and cloud resources. HPC manager is based on a queuing approach. The manager of the cloud resources is based on the REpresentational State Transfer (REST) Application Programming Interface (API). Both types of resources are often enriched by support from high-throughput resources or accelerated resources.

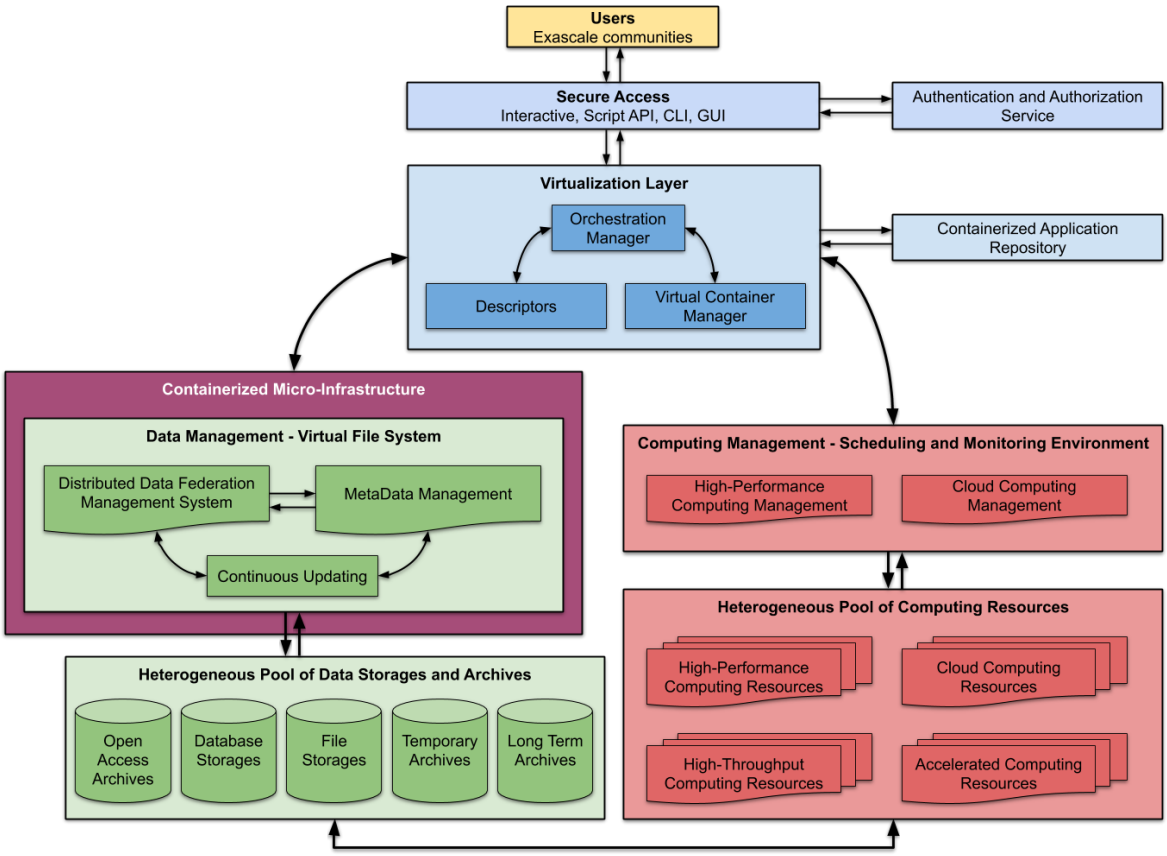

Figure 2:Reference exascale architecture.

4. Results

Experiments were computed on UvA (University of Amsterdam) and AGH (Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie) computing sites for Layer I and Layer II of the use case software. The use case application handling of the resources and access to CPU and GPU memory was intensively optimized. Table 2 and Table 3 illustrate the current status of the application layers I and II. First order statistics of computational requirements for Layer I are reported in Table 2. First order statistics for the time requirements of Layer II are reported in Table 3. Performance of Layer II is also reported in Table 3 in the form of model accuracy.

Table 2: Time baselines of the first layer: Data Preprocessing and Patch Extraction.

Location	Hes-so	UvA	AGH
Patch sampling time [s/patch]	0.41	-	-
Data loading time [ms/patch]	2.0	1.5	0.5

Table 3: Performance baselines for model 1 (ResNet50) of the second layer: Interchangeable Network Model.

Resource (Location)	Hes-so	UvA	AGH
Training accuracy	96,91±0,45	96,1±0,24	84.3
Validation accuracy	85,6±6,2	83,1±0,14	93.7
Training time [s/epoch]	2440,80	1203,68	17277
10 epochs time [h]	7 h	3.34 h	47 h

The performance evaluation highlights very efficient data extraction and loading at AGH, with only 0.5 ms to load a single patch. The high-performance GPUs available at UVA, by contrast, provide fast computations on a single GPU, halving the computational time of the model training (from 7 hours to 3.34 hours).

5. Conclusion

The paper presents the medical use case of the PROCESS project that is focused on exascale learning on medical image data. The requirements coming from the use case are handled by the exascale reference architecture. It is based on containerization and virtual machines supported by an exascale capable distributed virtual file system, and computing manager. The last section presents experimental results that were performed at two computing sites and show their advantages and disadvantages within the use case scenario.

Acknowledgement
This work is supported by the "PROviding Computing solutions for ExaScale ChallengeS" (PROCESS) project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777533, the project APVV-17-0619 (U-COMP) "Urgent Computing for Exascale Data", and the project VEGA 2/0167/16 "Methods and algorithms for the semantic processing of Big Data in distributed computing environment"

Martin Bobák (Dr.) is a research scientist at the Institute of Informatics of the Slovak Academy of Sciences. He received Ph.D. degree in Applied Informatics from the Institute of Informatics of the Slovak Academy of Sciences in 2017, and M.Sc. degree in Computer Science from Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava (Slovakia) in 2013. His research interests are cloud computing, algorithms and data structures. He is (co-)author of several scientific papers and has participated in international (European Union's Horizon 2020 research and innovation programme), and national research projects as a key person and scientific coordinator. He is a reviewer for international scientific conferences and journals, and a teaching assistant at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava (Slovakia).

Ladislav Hluchý (Assoc. Prof.) is the Head of the Parallel and Distributed Information Processing department, the vice-director of the Institute of Informatics, Slovak Academy of Sciences (IISAS). He received M.Sc. and Ph.D. degrees, both in computer science. He is RD Project Manager, Work-package leader in a number of 4FP, 5FP and 6FP projects, as well as in Slovak RD projects (VEGA, APVT, SPVV). He is a member of IEEE, ERCIM, SRCIM, and EuroMicro consortiums, the editor-in-chief of the journal Computing and Informatics. He is also (co-)author of scientific books and numerous scientific papers, contributions and invited lectures at international scientific conferences and workshops. He also gives lectures at Slovak University of Technology and is supervisor and consultant for Ph.D., master and bachelor studies.

Mara Graziani (M.Phil.) is currently a PhD student at the Computer Science faculty at the University of Geneva (Switzerland). Her research focus is on interpreting Deep Learning for medical applications. She completed the Masters of Philosophy in Machine Learning, Speech and Language Technology at the University of Cambridge (UK) in 2017. She has a BEng. in Information Technology Engineering at La Sapienza Univerisity of Rome.

Henning Müller (Prof.) studied medical informatics at the University of Heidelberg, Germany, then worked at Daimler-Benz research in Portland, OR, USA. From 1998-2002 he worked on his PhD degree at the University of Geneva, Switzerland with a research stay at Monash University, Melbourne, Australia. Since 2002, Henning has been working for the medical informatics service at the University Hospital of Geneva. Since 2007, he has been a full professor at the HES-SO Valais and since 2011; he is responsible for the eHealth unit of the school. Since 2014, he is also professor at the medical faculty of the University of Geneva. In 2015/2016 he was on sabbatical at the Martinos Center, part of Harvard Medical School in Boston, MA, USA to focus on research. Henning is coordinator of the ExaMode EU project, was coordinator of the Khresmoi EU project, scientific coordinator of the VISCERAL EU project and is initiator of the ImageCLEF benchmark that has run medical tasks since 2004. He has authored over 400 scientific papers with more than 12,000 citations and is in the editorial board of several journals.