Toward the convergence of in-situ data analysis and deep-learning methods for efficient post processing of pertinent structures among huge scientific datasets

S. Fiore[1], G. Aloisio[2], P. Ricoux[3], S. Brun[4], JM. Alimi[5], M. Bode[6], R. Apostolov[7], S. Requena[8]
[1]Euro-Mediterranean Center on Climate Change Foundation, Italy and ENES
[2]University of Salento, Italy
[3]Total, France
[4]CEA, France
[56]LUTh, CNRS, OBSPM, France
[6]RWTH Aachen University, Germany
[7]KTH and BioExcel CoE, Sweden
[8]GENCI, France

Introduction

The nature of science is changing – new scientific discoveries and socio-economical innovation are emerging from the analysis of large amounts of complex data generated by high-throughput scientific instruments (sequencers, synchrotrons, scanners, microscopes, …), observational systems (telescopes, satellites, network of sensors, …), extreme-scale computing (for both capability based large scale 3D simulations as well as ensemble or coupled multiscale/multiphysics simulations), and public World Wide Web.
In many domains – such as astronomy, physics, earth sciences, environmental sciences, genomics, biomolecular research, health sciences, financial, engineering, and social sciences, etc. – our ability to acquire and generate data is starting to outpace largely our ability to manage, explore, analyse, and valorise them both technically and socially, leading to the development of a new field called High Performance Data Analytics.
In order to be able to extract the wealth of information hidden in those data, and to valorise the infrastructures that generate them, new radical and holistic end-to-end data management approaches are needed.
Among many recent reports addressing such vision, the EESI2 project issued in 2015 toward the European Commission and national agencies several recommendations for the development of in-situ/in-transit post processing frameworks as well as identification of turbulent flow features in massively parallel Exascale simulations.
This position paper takes these recommendations as input and propose to increase their scope by adding machine/deep[1] learning capabilities for the development of real cognitive tools serving an accelerated science. It advocates urgently funding agencies for a joint call for proposals over 12 pilots, each cross-fertilizing experts of domain science and engineering (combustion, astrophysics, climate, life sciences, …), machine/deep learning as well as HPC experts and centres.

Key issues and scientific and industrial data analysis challenges

With regard to climate science, deep-learning techniques on large-scale datasets can provide a solid and advanced tool/methodology for understanding climate extremes (i.e. heat waves, tropical storms, and cyclones) detecting events, patterns, and trends as well studying their location, intensity, and frequency[2]. Challenges to face in this domain relate to the large amount of data to analyse, the different scales/resolutions, the complexity of the deep/machine learning algorithms and techniques integrated into a high performance scientific data management eco-system.

Numerical simulations play a fundamental role in cosmology today in the understanding of the origin and nature of the dark components of the universe as dark matter and dark energy and their influence on the

[1] Machine learning: a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of programs that can teach themselves to grow and change when exposed to new data. Deep learning: a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-linear transformations.
[2] http://www.wcrp-climate.org/grand-challenges/gc-extreme-events

formation of cosmic structures. In recent years, the first numerical simulations[3] of the full observable universe with dark energy were performed generating huge amounts of data (> PB) that must be analysed and valorised rapidly. But the understanding of processes in cosmology requires to study the structure and dynamics of numerous physical observables as the velocity fields, the gravitational potential, the deflection gravitational field, the gravitational wave field, ... the use of outcome of deep learning technology tool would allow high progress, but will require strong development of new techniques to be able to handle large volumes of multi-dimensional data.

In industry and especially in oil & gas (reservoir simulation) or in high fidelity combustion/multiphase simulations applied to turbines or engines, the need of in-situ data processing is already a current and critical problem at the Petascale era for several applications with huge amounts of raw data to manage. Typical data rates in the order of 350 TB/run in reservoir modelling or 1PB/30 min wall clock time in high fidelity combustion/multiphase will not even allow for the analysis of dynamic behaviour.
Current "on the fly" standard post-processing and analysis tools with features extraction applied to reservoir modelling or combustion/multiphase simulations are based on the use of different methods linking statistics, volume rendering, and topological data analysis. In that case coupling methods with new development of learning methods are certainly promising, in particular topology methods and the hierarchical deep learning[4]. They could be for example used to find correlations between the evolution of single droplets and turbulent length scales, which is still a challenge in high-fidelity multiphase simulations. A first joint project[5] between RWTH Aachen University and FZ Jülich, which focused on the development of large-scale smart in-situ visualization/post-processing techniques in the field of multiphase simulations, showed significant improvement in terms of I/O speed and data handling for Petascale simulations.

Computational study and design of molecules and materials on the atomistic scale is essential in the chemical, pharmaceutical and materials sciences and industries. It requires rigorous, unbiased, and accurate theoretical treatment. While numerical approximations to the many-electron problem are available, their enormous computational cost requires HPC to overcome current limits in terms of system size, simulation length, and size of databases in high-throughput screening. Concomitant with the increasing availability of big databases for chemical compounds, crystal structures, and now electronic structure calculations themselves, machine learning models are being developed[6] that interpolate between ab initio electronic structure calculations to accurately predict properties of new similar systems or to analyse high-volume data from simulations to uncover "hidden correlations" to gain new physical insights[7].

Finally, in the area of Life Sciences research, molecular simulation is a powerful tool to gain understanding of the structure and dynamic function of the basic building blocks of living organisms such as proteins, DNA, lipids, small molecules and up to the level of single cells. High-end compute infrastructures and highly scalable and efficient software packages are already capable of generating immense amounts of data e.g. by performing long time-scale simulations of multi-million particle systems or massive ensemble simulations of medium sized ones. Similarly, cryo-electron microscopy (cryo-em) methods have improved tremendously and are capable of elucidating the structure of large macromolecular complexes but their efficiency depends on the fast analysis of the terabytes of image data produced daily by the microscopes. Yet the tools for post-processing and analysis still lag in capabilities. Fostering research in the area will require development of the necessary software stack for exploitation of deep learning methods for analysis of multi-dimensional data in particular relevant to areas such as computational drug design, protein structure and function, cryo-em image processing etc.

[3] www.deus-consortium.org
[4] http://www-pequan.lip6.fr/~tierny
[5] M. Bode, J.H. Göbbert, H. Pitsch, "Detailed Investigation of Liquid Sheet Breakup Using Direct Numerical Simulation and In-situ Visualization", JARA|HPC project, 11/2014-10/2016.
[6] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, O. Anatole von Lilienfeld: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Physical Review Letters, 108(5): 058301, 2012. DOI 10.1103/PhysRevLett.108.058301
[7] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, Matthias Scheffler: Big Data of Materials Science: Critical Role of the Descriptor, Physical Review Letters, 114(10): 105503, 2015. DOI 10.1103/PhysRevLett.114.105503

Toward the convergence of scientific data analysis and machine/deep learning techniques

In 2015, inside the set of recommendations[8] issues by the EESI2 project, one-third were related to smart data analysis of scientific applications. Based on the rationale that the deluge of data generated by large scale or ensemble/coupled multiphysics/multiscale simulations become impossible in a competitive time to process by current techniques, the recommendations were proposing to develop at the European scale frameworks for in-situ/in-transit data analysis as well as providing to such tools the possibility to identify on the fly pertinent (turbulent) structures.
In-situ/transit technique allows to benefit from data locality, over the different memory hierarchies, just after the data is computed for performing real-time and non-intrusive post-processing of the raw data and thus reduce I/O overheads and optimize energy by storing only refined data. Such efficient post-processing could lead in as reverse loop to a new class of efficient computational steering techniques, again able to reduce both time and energy to solution.
Implementing such on-the-fly post processing tools need also to be able to implement high-order low pass and high pass filters, data mining features of reduction, cross-correlation, pattern/structure/field lines reconstruction/detection, ordering, partitioning, compression of data as well as trajectory based flow feature tracking.

Artificial intelligence methods become more widely adopted since the 2000s with machine learning techniques driven by rise of big data, and more recently by deep learning techniques driven by massively parallel computational hardware and new algorithms using multi-layer neural networks. Such models by using successive computational layers that process data in a hierarchical fashion by applying on each step convolutional layers (filters) have been widely adopted in image, video, sound and speech processing. Scientific communities as well as companies like Google, Amazon, Facebook, nVIDIA, Microsoft … developed machine learning frameworks like Torch, SPARK, Mahout, TensorFlow, DMTK, Shogun, Caffe, Theano, Scikit-Learn, that now start to be used as well in Life Sciences[9] or particle physics[10].

In the context of the BDEC conference about “Pathways to convergence” and the work conveyed by the application working groups of the EXDCI[11] H2020 European project, it is asked to the European Commission and the national funding bodies to organise a joint call for proposal toward the convergence of scientific data analysis and machine/deep learning.
By benefiting from the expertise of European teams in these fields, the objectives of this call could be to bridge on up to 12 pilot projects from many scientific and industrial fields the skills of experts in domain science, deep/learning and HPC experts. This mandatory cross-fertilisation of expertise could allow concretely during 3 years for each project to develop modern in-situ/in-transit post-processing techniques and assess the potential of machine/deep learning techniques for pertinent features detection in turbulent fluids, seismic processing, medical imaging, …

[8] http://www.eesi-project.eu/ressources/documentation/#eesi2-deliverables
[9] https://followthedata.wordpress.com/2015/12/21/list-of-deep-learning-implementations-in-biology/
[10] The Higgs ML challenge : https://higgsml.lal.in2p3.fr
[11] https://exdci.eu