Ph.D. Grant Application: Laboratory of Excellence UCN@SOPHIA

"Towards Green Learning applied to medical diagnostic" ou "Towards a Green Intelligence applied to health care and well-being"

Labex scientific themes: Big Scientific Data

Supervisors:

Michel RIVEILL (PU UNS – I3S), Frédéric PRECIOSO (PU UNS – I3S), Pascal STACCINI (PU-PH UNS INSERM UMR 912)

Context and Bottlenecks

"Big Data" and "Green" may sound as two of the top buzzwords at the moment like "Cloud" or "Fog". These words do not pop up from nowhere and are in fact related to the same phenomenon: the incredible increase of the Digital Universe. To the best of our knowledge, the term Digital Universe was first used, in 2007, by the International Data Corporation (IDC) in their first White Paper on forecasting both the amount of digital bits created, captured and replicated in the "Universe", and the possibilities provided by such an amount of information. Their forecasting estimation is an extension of the seminal work conducted in 2003 by Lyman and Varian, professors at the University of California at Berkeley (USA), in their study "How much information?" [1]. According to the IDC annual white papers, the Digital Universe was about 281 Exabytes (i.e. 281 Billions of Gigabytes) in 2007 [2] ; since then, it increased by about 6 times, cracking the Zettabyte barrier in 2010 (i.e. more than 1000 Exabytes) to pass 1.8 Zettabyte in 2011 (see [3] and [4]). Healthcare is one of the obvious examples of new segments of big data growth.

However, as pointed out by the IDC in [3] and [4], the vast majority of data in the Digital Universe is unstructured that means that most of the time, we know little about the data, unless it is somehow characterized or tagged. Building, extracting, collecting, and exploiting metadata is one of the fastest-growing sub-segments of the digital universe (though metadata itself is a small part of the digital universe overall). IDC experts believe that by 2020, a third of the data in the digital universe (more than 13,000 exabytes) will have Big Data value, but only if it is tagged and analyzed. Hence, designing methods to extract data characterization and content semantic description from unstructured content that can lead to information representations directly compatible with structured data and all powerful indexing techniques already emerging is one of the main Big Data challenges.

The last decade has seen the rise of machine learning methods for decision support and information retrieval in large data volumes. Successively, Boosting, Support Vector Machines, Random Forest, have built the success stories. More recently, the consistent efforts in artificial intelligence of the giants of the IT industry, such as Google, Microsoft, Yahoo, IBM, Facebook, have leaded to bring out models of artificial neural networks called "deep architectures". These deep learning techniques provide solution to high-stake applications such as information retrieval in huge volumes of multimedia content [5], as speech recognition [6, 7, 8], extraction of social communities from social networks [9], or knowledge extraction from text data [10].

However, one of the main drawbacks of deep architectures lies in the size of the training sets and the computation time required to train these networks. Although several open-source

implementations operating performances from GPU programming, the resources required to learn the right parameters of these architectures are considerable. As examples, the approach by Convolutional Neural Network (CNN) of Collobert et al. [10] using 852 million words and four weeks on one GPU computing to achieve the best performance of the state of the art in automatic language processing tasks and a similar architecture based on CNN proposed by Krizhevsky et al. [11] for classification of images in the largest image dataset available today, ImageNet, required the calculation of 60 million parameters with training process on two GPUs during a week. The most recent work on the basis ImageNet [12] leads to a CNN of 19 layers, for 144 million of parameters to be determined during the training phase. The training of such a network has required two to three weeks of calculations on a system with 4 GPU Titan Black NVidia (i.e. latest generation). Other more sophisticated methods still rely on parallel processing both for training the neural network and for distributing data by differentiating processing of the various layers of the network [13].

Even though very effective, such algorithms require always more expensive systems and always more energy which is not affordable in all contexts. This has led to the recent emergence of research efforts in the design of embedded low-power massive parallel processing machines [14].

Common understanding of “Well-Aging” research is often monitoring elder people with sensors so that their health status is permanently controlled (control of blood pressure, cardiac rhythm, or even alert and accident detection from home “videosurveillance”) so that they can both remain in their personal environment and be efficiently rescued in case of emergency. However, a less visible aspect of “Well-aging” research while being probably the biggest medical effort lies in the analysis of all data gathered from the appearance of a disease, until a possible hospitalization, including doctor visit and drug prescription reporting [15, 16, 17]: One very common example of this Well-aging research is the yearly intense observation process to detect influenza appearance every winter and the launch of adequate vaccination campaigns for the weakest ones (young and elder persons).

Current studies in this field are facing several obstacles: if privacy respect seems legally well-defines, practices managed by the medical services can’t achieve full standardization of the anonymization process, resulting most of the time in an impossible interoperability of the data collected by different public or private hospitals. Currently in order to get this level of analysis of the health situation in a given area, most of the effort is invested into back office ontology (i.e. ontology definition, alignment…) while the statistical analysis is mainly extracting correlations and co-occurrences.

The project in this thesis is to bridge the gap between very powerful software platform research, machine learning latest research and well-aging/well-being research.

Scientific objectives

Our main scientific objectives are:

1. Design new machine learning algorithms for embedded low power massive parallel processing machines
2. Design new distributed architectures based on these low power platform to improve learning results and make such “nano-supercomputer” architectures affordable and manageable for medical departments.

3. Design new learning algorithms for pushing well-aging and well-being studies a step further by providing big data oriented software solutions.

The health community is particularly interested in low-consuming solutions with analysis capabilities to help with diagnoses.

Complementary of teams involved in the project

The I3S SPARKS team has recognized expertise in the field of machine learning and in the software architecture with broad scales and the research team at University Hospital of Nice (Nice CHU) is a specialist for medical information systems and public health surveillance.

This project brings together 3 scientific communities represented each by one supervisor of this PD thesis:

- F. Precioso is Professor at Nice University and co-leads the Knowledge, Learning and Extraction activity in SPARKS Team of I3S Laboratory. He works on Machine Learning (ML) for computer vision problems, focusing on large-scale multi-class image database classification and interactive retrieval, on new low complexity machine learning methods for embedded systems and on methods learning jointly from unstructured (signal, images and videos) and structured data (metadata) for medical and biological domains.
- M. Riveill, is Professor at Nice University and co-leader of the Scalable Software Systems activity in SPARKS. His interests are design and analysis of software architecture for scalability problems.
- P. Staccini, is both Professor at Nice University and Public Health Practitioner in biostatistics, medical informatics and information technologies at the Faculty of Medicine of University of Nice-Sophia Antipolis. He is the head of the Department of Medical Informatics and Information of the Nice University Hospital, in charge of the the DRG-based accounting information system and medical activity assessment.

To validate our proposal we will use especially the health databases built for inpatients and outpatients care reports by Prof. P. Staccini and his group from University Hospital of Nice (Nice CHU). These databases, evolving annually, concern all institutions and health professionals. At a regional level, thanks to the resources from the CHU team in Nice, we will have access to the database of public and private hospital visits in PACA which represents more than a million of rows per year. At a national level, the data collected is about 20 billion rows of 30 months and we can access the last 10 years of collections of the University Hospital of Nice. Accessing to these data should enable us to compare approaches and validate the choices that must be made.

Moreover contacts have been made with NVidia France and teams of reference centers to benefit from the latest possibilities in terms of platforms.

These combined skills ensure access to medical data to enable validation of proposed approaches both in the design of learning algorithms and in the analysis of software requirements to enable proper management of the complexity and the size of data to be processed.

Relevance with the Labex axes and priorities

This project is relevant for the axes:

- Energy Efficiency: “To increase network energy efficiency, both for societal and personal reasons”
- Big Scientific Data
- eHealth (even though it is not “homecare services for persons with reduced autonomy” per se)

If the link with the first three themes seems obvious in view of the project, the link with the theme health deserves to be further explained.

The "health" databases we will mainly use in this thesis are those formed in the DRG-based medical reporting of both inpatients and outpatients.

One of the desired outcomes of the work that we will perform is to complete the studies that have begun on health medico-administrative data. These studies are limited for two reasons:

1. the understanding of encoded data into different databases from different hospitals,
2. Health data, even though anonymized, cannot be easily exploited “en masse” owing to a current lack of standardization of the anonymization process.

By designing new distributed architectures based on embedded low-power massive parallel processing machines, we will be able to deploy them locally on each site and thus avoid a heavy process of aligning the metadata between the resources. Furthermore, the efficiency of our solutions will allow us to address learning jointly from records of patient examinations and metadata.

Bibliography

[1] P. Lyman and H. R. Varian, "How much information?," The journal of electronic publishing, http://dx.doi.org/10.3998/3336451.0006.204, 2003.

[2] J. Gantz and D. Reinsel, "The Expanding Digital Universe," IDC white paper, http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf, 2007.

[3] J. Gantz and D. Reinsel, "Extracting Value from Chaos," IDC white paper, http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf, 2011.

[4] J. Gantz and D. Reinsel, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC white paper, http://idcdocserv.com/1414, 2012.

[5] Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, & A.Y. Ng, “Building high-level features using large scale unsupervised learning”, In International Conference on Machine Learning (ICML), 2012.

[6] A. Mohamed, G.E. Dahl & G.E. Hinton, “Acoustic Modeling Using Deep Belief Networks”, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp.14-22, 2012.

[7] G.E. Dahl, Y. Dong, L. Deng, A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”, in Audio, Speech, and Language Processing, IEEE Transactions on, 20:(1), pp.30-42, Jan. 2012.

[8] N. Jaitly, P. Nguyen, A. Senior & V. Vanhoucke, “Application of pretrained Deep Neural Networks to large vocabulary speech recognition”, in Proceedings of Interspeech, 2012.

[9] X. Glorot, A. Bordes & Y. Bengio, “Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach”, in International Conference on Machine Learning (ICML), vol. 27, pp. 97-110, 2011.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch”, in the Journal of Machine Learning Research, 12:2493--2537, 2011.

[11] A. Krizhevsky, I. Sutskever and G.E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, in Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[12] K. Simonyan, and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, in International Conference on Learning Representations (ICLR) Workshop 2015.

[13] A. Krizhevsky, “One weird trick for parallelizing convolutional neural networks”, in CoRR, abs/1404.5997, 2014.

[14] B. Hernandez, H. Perez, R. Isaac, S. Ruiz, O. DeGyves, and L. Toledo, L., “Simulating and Visualizing Real-Time Crowds on GPU Clusters”, in Computacion y Sistemas, vol. 18. pp. 651--664, 2014.

[15] R. Hamad, S. Modrek, J. Kubo, B.A. Goldstein, M.R. Cullen, “Using "big data" to capture overall health status: properties and predictive value of a claims-based health risk score”, in PLoS One. 2015 May 7;10(5), 2015.

[16] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, “Big Data Analytics in Healthcare,” in BioMed Research International, vol. 2015, 2015.

[17] Almalki, Manal, Kathleen Gray, and Fernando Martin Sanchez. “The Use of Self-Quantification Systems for Personal Health Information: Big Data Management Activities and Prospects.” Health Information Science and Systems 3.Suppl 1 (2015): S1. PMC. Web. 11 Sept. 2015.