Using MapReduce for Large–scale Medical Image Analysis

Dimitrios Markonis, Roger Schaer, Ivan Eggel, Henning Müller, Adrien Depeursinge
University of Applied Sciences Western Switzerland (HES–SO), Business Information Systems, Sierre, Switzerland

***Index Terms*—large–scale; medical; image analysis; big data; scalability; MapReduce; Hadoop; support vector machines; content–based image retrieval; texture analysis;**

I. Introduction

The growth of the amount of medical image data produced on a daily basis in modern hospitals forces the adaptation of traditional medical image analysis and indexing approaches towards scalable solutions [1]. The number of images and their dimensionality increased dramatically during the past 20 years. Recent progress in image processing and machine learning makes it possible to assist clinicians in the detection and characterization of important events in large image series. However, the process of extracting intricate features from large datasets of 3D/4D images, as well as training machine learning algorithms and global system optimization are extremely demanding in terms of computation time, storage capacity and network bandwidth [2]. The MapReduce framework is a distributed computing framework and has recently been used for large–scale image description and analysis [3]. In this work, MapReduce is used to speed up and make possible three large–scale medical image processing use–cases: (i) parameter optimization for lung texture classification using support vector machines (SVM), (ii) content–based medical image indexing, and (iii) three–dimensional directional wavelet analysis for solid texture classification.

II. Methods

A cluster of heterogeneous computing nodes was set up in our institution using Hadoop allowing for a maximum of 42 concurrent map tasks. The majority of the machines used are desktop computers that are also used for regular office work. A minimum of two logical cores were not allocated to the Hadoop TaskTracker process, ensuring that common daily tasks could still be run smoothly.

III. Results

Parallel grid search for optimal SVM parameters[1] was carried out on the Hadoop cluster. A map task was defined for each coupled value of (C, σ). A clear link between the runtime of a map task and the resulting classification accuracy was observed: most of the tasks with long runtimes resulted in poor classification accuracies. The interruption of such map tasks allowed a reduction of the total runtime from 50h to 9h15m, while keeping all coupled values (C, σ) leading to best classification performance. A sequential execution would require 990h approximatively on a desktop computer.

Two approaches for content–based image indexing were compared and implemented in the MapReduce framework: component–based versus monolithic indexing. The former is convenient to separately optimize feature extraction and the indexer because it does not require to run the whole pipeline for each optimization. However, it requires to write the features to a very large CSV (Comma–Separated Values) file of approximately 100 Gb for 100,000 images. This resulted in an unexpectedly long runtime for the feature extractor with the MapReduce framework in the component–based approach. The result is consistent with previous work that showed that MapReduce was not performing well with input-output (IO)–intensive tasks [4]. The monolithic strategy showed to be well–suited for MapReduce, which allowed indexing 100,000 images in about one hour using 24 concurrent tasks.

The parallelization of solid texture processing based on non–separable three–dimensional Riesz wavelets allowed to reduce a total runtime from more than 130h to less than 6h, while keeping the code in the original Matlab/Octave programming language with Hadoop streaming.

IV. Discussions and Conclusions

Overall Hadoop has shown its utility for large scale medical image computing. The three use–cases reflect the various challenges of processing medical visual information in clinical routine: parameter optimization, indexation of image collections with hundreds of thousands images, and multi–dimensional medical data processing. In all tasks very positive results could be obtained helping the projects to scale with limited local resources available and moderate efforts to adapt the software.

[1] SVM parameters are the cost C of the variance σ of the Gaussian kernel.

References

[1] K. P. Andriole, J. M. Wolfe, and R. Khorasani. Optimizing analysis, visualization and navigation of large image data sets: One 5000–section CT scan can ruin your whole day. *Radiology*, 259(2):346–362, May 2011.

[2] U. Catalyurek, S. Hastings, K. Huang, V. S. Kumar, T. Kurc, S. Langella, S. Narayanan, S. Oster, T. Pan, B. Rutt, X. Zhang, and J. Saltz. Supporting large scale medical and scientific datasets. In *Parallel Computing: Current & Future Issues of High–End Computing, Proceedings of the International Conference ParCo 2005*, pages 3–14, 2005.

[3] B. White, T. Yeh, J. Lin, and L. Davis. Web–scale computer vision using MapReduce for multimedia data mining. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, July 2010.

[4] S. Loebman, D. Nunley, K. Yong-Chul, B. Howe, M. Balazinska, and J. P. Gardner. Analyzing massive astrophysical datasets: Can Pig/Hadoop or a relational DBMS help? In *IEEE International Conference on Cluster Computing and Workshops*, pages 1–10, August 2009.