Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis

Allan Hanbury[1], Henning Müller[2], Georg Langs[3], Marc André Weber[4], Bjoern H. Menze[5], and Tomas Salas Fernandez[6]

[1]Vienna University of Technology, Austria
[2]University of Applied Sciences Western Switzerland (HES–SO), Switzerland
[3]CIR, Dep. of Radiology, Medical University of Vienna, Austria
[4]Radiology Department, University of Heidelberg, Germany
[5]ETHZ, Zürich, Switzerland
[6]Gencat, Spain
henning.mueller@hevs.ch

Abstract. Benchmarks have shown to be an important tool to advance science in the fields of information analysis and retrieval. Problems of running benchmarks include obtaining large amounts of data, annotating it and then distributing it to the participants of a benchmark. Distribution of the data to participants is currently mostly done via data download that can take hours for large data sets and in countries with slow Internet connections even days. Sending physical hard disks was also used for distributing very large scale data sets (for example by TRECvid) but also this becomes infeasible if the data sets reach sizes of 5–10 TB. With cloud computing it is possible to make very large data sets available in a central place with limited costs. Instead of distributing the data to the participants, the participants can compute their algorithms on virtual machines of the cloud providers. This text presents reflections and ideas of a concrete project on using cloud–based benchmarking paradigms for medical image analysis and retrieval. It is planned to run two evaluation campaigns in 2013 and 2014 using the proposed technology.

Keywords: benchmark, medical image analysis, anatomy detection, case–based medical information retrieval, cloud computing

1 Introduction

In many scientific domains benchmarks have shown to improve progress, from text retrieval (TREC, Text Retrieval Conference [4]), to video retrieval (TRECvid, TREC video task [7]), image retrieval (ImageCLEF, image retrieval track of the Cross Language Evaluation Forum, CLEF [5]) and object recognition (PASCAL [3]). Medical applications have also been subject to benchmarks such as ImageCLEFmed on visual data, to text retrieval from patient records in TREC. Impact of the benchmarks was shown in [6, 8, 9], both economically and scholarly.

Data, particularly visual data, has been difficult to obtain for many years and thus data sets used for evaluation have often been small as a result. With the

creation of social data sharing sites such as YouTube[1] and FlickR[2], obtaining large data sets has become much easier as many images are made accessible with clear licenses for their use, most often using Creative Commons licenses. In the medical field the funding agencies also push for open data accessibility and this means that data have now become available on a larger scale. Getting terabytes of data is in principle no longer a major difficulty.

The problem has rather become the annotation or ground truthing of large amounts of existing data that is often very expensive. In the case of medical data the ground truthing most often needs to be performed by experts, leading to even higher costs. Expert judgements are also a limitation for crowd sourcing approaches [1] that can otherwise help limiting costs for relevance judgements.

This text proposes solutions for the data distribution challenge by using an infrastructure based on cloud computing [2]. Bringing the algorithms to the data may allow for a better comparability of approaches, and it may make it better possible to work on sometimes restricted data. Virtual machines in the cloud that have access to the data allow all participants to use their choice of operating system and environment. Making code work in a different run time environment can sometimes be a tedious task and it can also limit participation. Having a similar virtual machine for each participant also creates the same conditions for all participants in terms of processing speed and optimization. In many standard benchmarks, the groups with a larger server capacity often have an easier task when trying to obtain very good results and test varying parameters.

Also the problem of ground truthing is tackled by the approach described in this paper, using a small gold (manually labelled) and then a large silver (fusion of participant submissions) ground truth set. Such a silver ground truth is planned to be generated through the results of the participants' runs in the cloud and can thus be created directly with the data and the algorithms. Putting such a ground truth together may lead to better analysis of techniques but there are also risks that the techniques of existing systems could bias the results in a similar way that pooling does.

This text also reflects on related ideas such as continuous evaluation when data remains available over a long term. Sharing environments might also help participants to collaborate and develop tools together and thus it can be a first step to facilitating component–based evaluation.

2 Materials and methods

This article is mainly based on reflections of how to leverage visual medical image analysis and retrieval to a new scale of processing, starting with simpler tasks (such as anatomy detection) and very large amounts of medical data (on the order of 10 TB), and then moving toward more complex tasks such as the retrieval of similar cases. All authors reflected on the topic to develop a new benchmark on medical visual data analysis and retrieval. The outcomes are based

[1] http://www.youtube.com/
[2] http://www.flickr.com/

on all constraints of the system, such as very large scale processing and the requirement to generate ground truth with expert involvement. The results of this are planned to be implemented in an EU funded effort named VISCERAL[3] (VISual Concept Extraction challenge in RAdioLogy). This paper only includes reflections and only few experiences with the described methods. Experience from the first setup and evaluation sessions are planned to follow.

3 Results

This section describes the main ideas based on the reflections on requirements of a benchmark for such a large amount of data that require expert annotations.

3.1 Infrastructure considerations

In terms of data distribution it is clear that going beyond several terabytes requires most research groups to change current infrastructures. Not only hard disks are required for this but also redundancy in case the disks fail, and quick access to the data to allow for processing in a reasonable amount of time. Cloud computing has the advantage that redundancy and backups are dealt with by the provider and not by the researchers. Access to the data can be given without the requirement to download the data and store them locally, so at any given time only part of the data is being treated making all data handling much easier for participants and organizers of such as challenge. The data can also be controlled better, meaning that confidential data can be used by the virtual machines and each use of the data can be logged, avoiding uncontrolled distribution. Participants can of course download small training data sets to optimize algorithms locally and then install the virtual machines for their specific setup, and run their algorithms on the cloud accessing the training data. This concept is also detailed in Figure 1. Execution will thus be in standard environments, allowing the evaluation of the efficiency of the tools, while groups with extremely large computing resources will not have major advantages.

The execution of the benchmark could then be done by the organizers by simply changing the path to the data in the tools of the participants and running the tools on the full data as shown in Figure 2. This has the advantage that 'cheating" or manual parameter tuning on the test data can be excluded as participants do not have access to the test data to use it for optimizations.

Such an infrastructure could also foster collaborations as systems can make services for specific tasks available easily and thus share components with other participants. This can help when some groups are specialized in text retrieval and others in visual image retrieval, for example. When the data can be made available long term, such an approach can also help creating a continuous evaluation where all groups using the data at later stages can submit their results via a standard interface. The algorithms can then be compared for efficiency, and bias towards groups with much computing power can be avoided.

[3] `http://www.visceral.eu/`

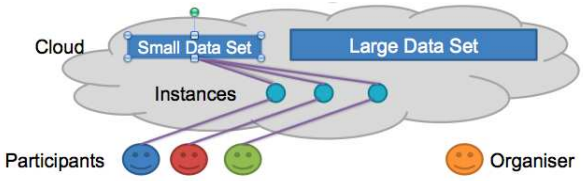

Fig. 1. The participants each have their own computing instance in the cloud, linked to a small dataset of the same structure as the large one. Software for carrying out the competition objectives is placed in the instances by the participants. The large data set is kept separate.

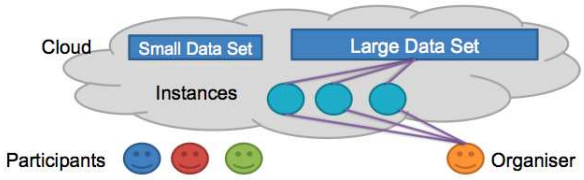

Fig. 2. On the competition deadline, the organiser takes over the instances containing the software written by the participants, upgrades their computing power, links them to the large data set, performs the calculations and evaluates the results.

3.2 Silver and gold corpora

Manual work is necessary to create high quality annotation. In the medical field this is expensive but essential for good evaluation. By outsourcing the work to countries with lower income the costs can be reduced but quality control is necessary, as errors can lead to meaningless evaluation results. Sharing results among many research groups as is the case in a competition also leads to much more efficient annotation as data is not only used in a single center. All manual annotation cannot scale to millions of images and some automation in the ground truth generation will be necessary to allow for scaling.

Using the results of all participants directly in the cloud to create a so–called silver corpus in addition to a manually annotated gold corpus can make it possible to compare results based on two data sets and analyze how well the performance measures compare. The silver corpus can be created as a majority vote of the results of all participant runs directly in the cloud. One of the risks is that many systems using similar techniques will dominate the silver corpus. It can however also be an option that part of the silver corpus, for example documents with disagreement, can be manually judged to estimate the number of errors or inconsistencies in the silver corpus. Albeit not an optimal solution, such ground truth can potentially increase data set size used for an evaluation and limit the resources necessary to create annotated data sets. This can make evaluation on extremely large data sets feasible, which would not be the case without automation.

3.3 Further reflections

Besides the purely technical reasons of allowing access to very large amounts of data there are several other aspects that could be improved by such a process. Research groups having less computing power are currently disadvantaged in evaluation campaigns. More complex visual features or data analysis can be extremely demanding in terms of computing power, so that many groups could simply not implement such complex approaches on large data. Measuring execution times has been proposed in the past but this is hard to control as the exact execution environment is rarely known. In terms of storage, currently few research groups would have the resources to process over 10 TB of data as not only the raw data but also computed data such as features need to be stored. Making available to participants the same types of virtual machines would give all groups the same starting point and full access to the data.

Another potential advantage of using a cloud–based approach is that public access can be limited to a training data set and then the virtual machines can be used to compute on potentially restricted data. This can for example be medical data, where anonymization can be hard to control as for free text but also intelligence or criminal data that cannot simply be distributed.

4 Discussion and conclusions

When organizing benchmarks using extremely large data sets, using the cloud seems the only possibility, as the algorithms need to be brought to the data rather then the other way around. In terms of pricing, the data transfer is actually a fairly expensive part and renting computing power less so. Bandwidth is also a problem in many other environments such as hospital picture archives or data distribution to participants in a benchmark. Such a system allows for a better comparison of techniques and creates equal possibilities for groups from all countries, with fewer disadvantages if weaker computing servers are available for optimization. This can also avoid using the test data for parameter tuning.

Silver corpora can strengthen the effect that standard techniques and not new approaches will be used, a typical criticism of benchmarks. Still, academic research needs to start using extremely large data sets as problems on big data are different from problems on smaller amounts of data. For discovering these challenges big data and large corpora are a requirement. Contradictions and confirmations can be found by comparing the results with the gold test corpus and the silver corpus and analyzing what precisely these differences might mean.

The mentioned data volumes will allow moving closer toward using the volumes commonly produced in hospitals, which is in the order of several terabytes per year. Simple pretreatment is required to make algorithms scalable including parallelization techniques such as Hadoop/MapReduce, used in web search. Still, most currently published research only uses very small data sets limiting the potential impact. Bringing the algorithms to the data and having research groups collaborate in the cloud on image analysis challenges will deliver new

research results and has the potential to bring medical image analysis one big step closer to clinical routine.

5 Acknowledgments

This work was supported by the EU in the FP7 through the VISCERAL (318068), PROMISE (258191) and Khresmoi (257528) projects.

References

1. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. ACM SIGIR Forum 42(2), 9–15 (2008)
2. Buyya, R., Yeo, C.S., Venugopal, S.: Market–oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In: 10th IEEE International Conference on High Performance Computing and Communications. pp. 5–13. IEEE (2008)
3. Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J.D.R., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J.: The 2005 PASCAL visual object classes challenge. In: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment (PASCAL Workshop 05). pp. 117–176. No. 3944 in Lecture Notes in Artificial Intelligence, Southampton, UK (2006)
4. Harman, D.: Overview of the first Text REtrieval Conference (TREC–1). In: Proceedings of the first Text REtrieval Conference (TREC–1). pp. 1–20. Washington DC, USA (1992)
5. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
6. Rowe, B.R., Wood, D.W., Link, A.N., Simoni, D.A.: Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. Tech. Rep. Project Number 0211875, RTI International (2010)
7. Smeaton, A.F., Kraaij, W., Over, P.: TRECVID 2003: An overview. In: Proceedings of the TRECVID 2003 conference (Dec 2003)
8. Thornley, C.V., Johnson, A.C., Smeaton, A.F., Lee, H.: The scholarly impact of TRECVid (2003–2009). JASIST 62(4), 613–627 (2011)
9. Tsikrika, T., Seco de Herrera, A.G., Müller, H.: Assessing the Scholarly Impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (Sep 2011)