Cloud–based Evaluation Framework for Big Data

Allan Hanbury[1], Henning Müller[2], Georg Langs[3], and Bjoern H. Menze[4]

[1] Institute of Software Technology and Interactive Systems,
Vienna University of Technology, Austria
hanbury@ifs.tuwien.ac.at
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
henning.mueller@hevs.ch
[3] CIR Lab, Department of Radiology, Medical University of Vienna, Austria
georg.langs@meduniwien.ac.at
[4] Computer Vision Laboratory, ETH Zürich, Switzerland
bjoern@ethz.ch

Abstract. The VISCERAL project is building a cloud-based evaluation framework for evaluating machine learning and information retrieval algorithms on large amounts of data. Instead of downloading data and running evaluations locally, the data will be centrally available on the cloud and algorithms to be evaluated will be programmed in computing instances on the cloud, effectively bringing the algorithms to the data. This approach allows evaluations to be performed on Terabytes of data without needing to consider the logistics of moving the data or storing the data on local infrastructure. After discussing the challenges of benchmarking on big data, the design of the VISCERAL system is presented, concentrating on the components for coordinating the participants in the benchmark and managing the ground truth creation. The first two benchmarks run on the VISCERAL framework will be on segmentation and retrieval of 3D medical images.

Keywords: Evaluation, Cloud Computing, Annotation, Information Retrieval, Machine Learning

1 Introduction

Demonstrating progress in data-centric areas of computational science, such as machine learning and information retrieval, requires demonstrating that a new algorithm performs better in its task than state-of-the-art algorithms. However, even though a continuous stream of published papers claim to have demonstrated such improvements, some scepticism remains. Hand [9] discusses the “illusion of progress” in classifier technology, while Armstrong et al. [4] present evidence for “improvements that don’t add up” in information retrieval (IR).

Evaluation campaigns and benchmarks aim at quantifying the state-of-the-art by making available tasks and data, and objectively comparing the results of multiple participants' approaches to performing the set tasks on the provided data. In the area of IR, evaluation campaigns have been run for over 20 years [10]. Current evaluation campaigns include TREC (Text REtrieval Conference)[5], TRECVid (TREC Video Evaluation)[6], CLEF (Cross Language Evaluation Forum)[7], ImageCLEF [14], NTCIR (NII Test Collection for IR Systems)[8], INEX (Initiative for the Evaluation of XML Retrieval)[9] and FIRE (Forum for Information Retrieval Evaluation)[10]. In the area of machine learning, the PASCAL challenges are well known[11], while in the area of medical image analysis, annual challenges are organised as part of the Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)[12].

However, even with these evaluation campaigns and challenges, a number of causes contribute to the above-mentioned lack of clear improvement:

Data: Even though evaluation campaigns, challenges and other mechanisms lead to the availability of commonly used test datasets "standardised" within communities, these datasets are often not a "realistic" approximation of real-world datasets. Reasons for this are that the datasets are often small so as to simplify dissemination and reduce computation time; are not representative of all of the the variation found in real-world data; or are cleaned in some way to reduce noise and outliers. Furthermore, it is unlikely that a single algorithm will have the best performance on all possible datasets. Concentration on achieving improvements on a few datasets could lead to algorithms highly optimised for these datasets. Alternatively, if many datasets are available, then results could be presented only on datasets for which performance is good. Finally, even though many such datasets are available, proprietary data are still often used in publications.

Algorithms: Source code or even executables for cutting edge algorithms are usually not made available. It is often difficult to re-implement an algorithm based on its description in a paper. This means that comparisons to such an algorithm can only reliably be made on the datasets on which it was tested in the original publication, and it is not possible to judge the performance of this algorithm on new data.

Baselines: New performance results in publications are often compared to a low and little optimized baseline, and not to the state-of-the-art algorithms. This is linked to a certain extent to the difficulty in obtaining or re-implementing state-of-the-art algorithms mentioned in the previous point. However, it could also be linked to the pressure to show some sort of improvement in order to get a paper published. Evaluation campaigns and challenges aim to

[5] http://trec.nist.gov/

[6] http://trecvid.nist.gov/

[7] http://www.clef-campaign.org/

[8] http://research.nii.ac.jp/ntcir/index-en.html

[9] https://inex.mmci.uni-saarland.de/

[10] http://www.isical.ac.in/~clia/

[11] http://pascallin2.ecs.soton.ac.uk/Challenges/

[12] http://www.grand-challenge.org has an overview of most MICCAI challenges.

solve this problem, but beyond the single publication incorporating the results of all algorithms submitted, in general little comparison to these results subsequently takes place.

Related to the above points, in the computational sciences in general, there has been recent concern expressed about the lack of reproducibility of experimental results, raising questions about their reliability. The lack of publication of program code has been identified as a significant reason for this [6]. There is currently work underway to counter this situation, ranging from presenting the case for open computer programs [12], through creating infrastructures to allow reproducible computational research [6] to considerations about the legal licensing and copyright frameworks for computational research [18].

Despite these shortcomings, evaluation campaigns do make a significant economic and scholarly impact. TREC, organised by the National Institutes of Standards and Technology (NIST) in the USA, is the longest running IR evaluation campaign and has been running since 1992. A 2010 study of the economic impact of TREC[13] came to the conclusion that "US$16 million of discounted investments made by NIST and others in TREC have resulted in US$81 million in discounted extrapolated benefits or a net present value of US$65 million". This is due to, amongst others, making available evaluation resources at a relatively low cost, developing and publishing evaluation methodologies, encouraging development of improved IR techniques and allowing companies to see which are the most successful techniques to integrate into their products. Recently, the assessments of the scholarly impact based on bibliometric measures have been published for two evaluation campaigns: ImageCLEF [20] and TRECVid [19]. Both papers demonstrate the impact through the high number of citations of papers written as a result of the evaluation campaigns.

The VISCERAL project[14] is developing a cloud-based framework for experimentation on large datasets, with a focus on image analysis and retrieval in the medical domain. Initially, it is aiming to reduce the complexities and barriers to running experiments on huge representative datasets, discussed in more detail in Section 2.

For a benchmark that is run on the VISCERAL framework, the task will be specified and the training data will be placed on the cloud. Participants will program solutions to the task in computing instances (virtual machines) on the cloud, effectively *bringing the algorithms to the data* instead of the more conventional transfer of the data to where the algorithms are. Benchmark organisers will then evaluate the task solutions on an unseen dataset. Over the next two years, two benchmarks for 3D medical imaging will be run: classification of regions of medical images and retrieval of medical images [13]. The benchmarks will be run on a dataset of at least 2TB of radiology images and associated radiology reports. This evaluation framework under development in VISCERAL is presented in Section 3. Finally, Section 4 discusses various considerations for further development of the evaluation framework. Further information on the

[13] `http://trec.nist.gov/pubs/2010.economic.impact.pdf`

[14] `http://visceral.eu`

VISCERAL project is available targeted at the IR community in [8], and at the medical imaging community in [13].

In other scientific fields such as physics, large Grid networks [5] such as EGI (European Grid Initiative) or previously EGEE (Enabling Grids for E-Science in Europe) [7] have been created for distributed data analysis where several large institutions share large infrastructures in virtual organizations. Still, most Grid middleware is hard to install and many small research groups in computer science do not have the funding to maintain such infrastructures, meaning that this model cannot be transferred to all fields of science and the medical imaging field is an area where this can be problematic [15]. A centralized infrastructure has the advantage of a very low entry level for groups to participate in such a campaign.

2 Challenges in Benchmarking on Big Data

The standard model used by the majority of evaluation campaigns in machine learning follows the following steps, where **O** indicates that the step is performed by the organisers and **P** indicates that the step is performed by the participants:

1. **O**: The organisers define the task to be performed and collect suitable data for the task. The data are divided into a training and test set. Sufficient ground truth for the task is created, where ground truth is required on the training data for training the machine learning algorithms, and required on the test data for evaluating the performance of algorithms by comparing their output to the test data ground truth.
2. **O**: The organisers publish the task information, and make the training data and associated ground truth available for download.
3. **P**: Participants train their algorithms using the training data and ground truth.
4. **O**: At a later date, the test data (without ground truth) is made available to the participants to download.
5. **P**: The participants run their trained algorithms on the test data, and submit the outputs (in a pre-defined format) to the organisers by a specified deadline (usually through an online submission system).
6. **O**: The organisers evaluate the performance of the algorithms on the test data using the corresponding ground truth, and release the results.

For IR benchmarks, the sequence is similar, except that the full set of data (often with some example queries and relevant documents) is released in step 2 for the participants to index in step 3. In step 4, the test queries are released, and the participants must submit the documents returned by each test query in step 5. While it would in theory be possible to provide the ground truth for the relevance of each document to the test queries in step 1, this would in practice require infeasible amounts of human input. In practice, the human input for the relevance judgements is provided in step 6, where relevance judgements are only done on documents returned by at least one algorithm, usually involving a technique such as pooling to further reduce the number of relevance judgements to be made [17].

For applications that in practice involve the processing and analysis of large amounts of data, running benchmarks of the algorithms on representative amounts of data has advantages. Using more data implies that the benchmark data can be more characteristic of the data used in practice, especially in terms of fine features of the data distribution and of the inclusion of outliers. However, running benchmarks on multiple terabytes of data leads to a number of practical problems, including:

Data distribution: Downloading the data through the internet can take an excessive amount of time and is costly. A solution that is often adopted is to send the data to the researchers on hard disks through the postal service. This however requires additional effort on the part of the benchmark organisers and involves higher costs.

Computing power: Not all researchers wishing to participate in the benchmark have sufficient local computing resources (processing and storage capacity) to participate effectively in the benchmark. For some research groups, processing all of the data potentially requires several weeks, while for others several hours may be sufficient.

Comparison of algorithm efficiency: As each participant runs their algorithms on a different infrastructure, comparison of the algorithm efficiency in terms of processing speed is not possible.

Obtaining sufficient ground truth: With more data, correspondingly more ground truth is necessary to make the use of the data advantageous. This means that manual annotation costs increase.

In IR, the largest benchmark datasets available are the ClueWeb datasets. The ClueWeb12 dataset[15] contains 870,043,929 English web pages, with an uncompressed size of 32 TB (5.3 TB compressed). It is distributed on hard drives sent by post.

The high cost of obtaining ground truth can be mitigated in some cases by the use of more cost-effective annotators, such as those available on crowdsourcing platforms [1, 21]. However, this approach is often not suitable for more specialised tasks requiring the annotators to possess expert knowledge.

3 VISCERAL Framework

The VISCERAL cloud-based benchmarking framework currently under development is shown in Figure 1. This represents the envisaged setup for a machine learning benchmark, but an IR benchmark would have a similar setup, with the main change being a single dataset available for indexing. The main components of the framework for managing the benchmark are the *Registration System*, which handles participant registration and management, and the *Analysis System*, which handles the running and analysis of results during the test phase, as

[15] `http://lemurproject.org/clueweb12.php`

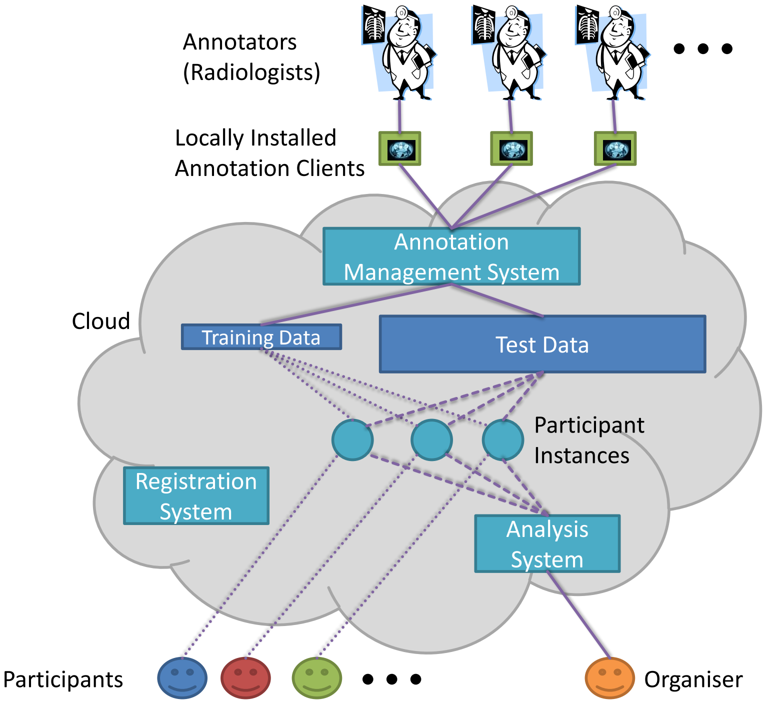

Fig. 1. VISCERAL evaluation framework. The dotted lines represent connections during the training phase, while the dashed lines represent connections during the test phase. Solid lines are connections that exist before and throughout the benchmark.

described in Section 3.2. The *Annotation Management System* coordinates the manual annotation, as described in Section 3.3.

3.1 Cloud-Based Framework

The cloud has innovated a number of aspects of computing, as it provides the appearance of infinite computing resources available on demand, eliminates upfront commitment by cloud users and provides the ability to pay for the use of computing resources on a short-term basis as needed [2]. The abilities necessary for the experimental approach described in this paper are:

- Provide the ability to centrally store and make available large datasets — Cloud providers already provide this service. For example, Amazon hosts public datasets free of charge[16].

[16] http://aws.amazon.com/publicdatasets/

- Allow multiple users to process the stored data without requiring the data to be transfered elsewhere — this is done through linking virtual storage drives to computing instances as required. For example, Amazon public datasets are accessed in this way.

The cloud-based framework allows the four practical problems of running benchmarks on big data listed in Section 2 to be overcome to a certain extent. The *data distribution* problem is solved by placing the data in the cloud and by having the participants install their software in computing instances in the cloud. All participants will have access to sufficient *computing power* in the cloud computing instance, and will have the choice of using a Linux or Windows instance. As the computing power is standardised, it will be possible to measure the *algorithm efficency* objectively. Finally, even though experts are required for creating the ground truth, the annotation process can be managed as a function of the inter-annotator agreement and participant entries to be efficient.

3.2 Benchmark Activity Flow

The benchmark consists of two phases, the *training phase* and the *test phase*. During the training phase, potential participants can register using the Registration System. During this registration, participants will be asked to sign a document regulating what can be done with the data. Once the signed document is uploaded, the benchmark organiser approves a participant. After the approval, participants are given access to a computing instance linked to the training data (indicated by the dotted lines in Figure 1). The participant has until the submission deadline to implement the necessary software in the computing instance to perform the benchmark task. The organisers will carefully specify parameters such as output format and naming of the executable files to allow automation of the calling of the programs. Participants will also have the possibility to download a subset of the training data, hence allowing optimisation to be done on local computing infrastructure if this is desired.

After the submission deadline, the Analysis System takes over control of all computing instances from the participants, and participants lose access to their instances. The computing instances are then all linked to the test dataset (indicated by the dashed lines in Figure 1). The Analysis System runs the software, analyses the outputs and computes the performance metrics. These performance metrics are then provided to the participants.

3.3 Manual Annotation

The use of the cloud allows the manual annotation of the data to be effectively controlled, which will be done by the Annotation Management System. As the VISCERAL benchmarks are using radiology image data, expert annotators in the form of radiologists will be used. As the time of such experts is expensive, it is important that the most effective use is made of their time.

The first annotation task, taking place before the begin of the training phase, is the annotation of the training data. The manual annotations will form the *gold corpus* used during the training phase. The radiologists performing the annotation will install a local client that assists in the manual marking of volumes of the images by using semi-automated techniques. The Annotation Management System will be able to assign images and anatomical structures to specific radiologists to annotate. It will assign part of the images to at least two radiologists so that inter-annotator agreement can be measured, for each annotated structure. This allows the system to measure the quality of the annotation, estimate the ability of the various annotators, and assign "difficult" images having lower inter-annotator agreement to more annotators of higher ability. Since only a part of the overall data will be annotated to form the gold-corpus the choice of cases to annotate is important. Based on the cumulative annotations, the system estimates the information gain expected from a particular case, and assigns those cases where this is maximal. This ensures that the variability represented in the gold corpus is large.

For evaluation of the participants' automatic annotation results on the test set, ground truth annotations are also necessary for this part of the data. Part of these annotations will be created by radiologists as described for the training set. However, due to the huge amount of data in the test set, manual annotation of all of it is infeasible. Therefore, a *silver corpus* approach, such as the one used in the CALBC challenges[17] [16], will be adopted. This silver corpus is built based on voting on the participant submissions. However, the Annotation Management System will also be able to request manual corrections of images for which it appears that the voting is inconclusive.

4 Discussion and Conclusion

The first two benchmarks organised in the VISCERAL framework will work on large scale 3D medical imaging. The first benchmark on segmentation of organs represents a machine learning style of evaluation with separate training and testing datasets, while the second benchmark on retrieval of similar images represents an IR style of evaluation. These first two benchmarks will be run in the style of a classic challenge or evaluation campaign, with a strict submission deadline for algorithms and resulting metrics of the evaluation being released by the organisers simultaneously. The aim is however to automate the process, allowing participants to submit algorithms in computing instances at any time, and to get rapid feedback about the calculated metrics directly from the system. The system could then also store results of all algorithms submitted, allowing effective comparison of a submitted algorithm with state-of-the-art algorithms. However, the willingness of researchers to use such a system allowing direct comparison to state-of-the-art algorithms must be investigated, as initiatives to provide such services have not been well accepted. For example, in the IR community, the EvaluatIR system [3] was hardly used, even though it provided

[17] `http://calbc.eu`

the choice of using it without being obliged to reveal the results obtained to other users.

For the initial two benchmarks run on this cloud-based evaluation framework, participants will have their cloud computing costs funded by project funds of the organisers. However, for sustainability of the framework, further models of financing the participants' computing costs will have to be developed. The simplest is that participants finance their own computing costs, although this will likely disadvantage groups with few financial resources. Alternatively, the cloud service providers could provide computing time to researchers in the form of grants (for example, it is currently possible for researchers to apply for grants in the form of free usage credits from Amazon[18]). Finally, a publicly-funded cloud-based evaluation infrastructure hosting standard datasets could provide subsidised or free access to researchers based on an application scheme.

An important consideration is who should provide this cloud-based evaluation service. Commercial cloud providers are already able to provide it, but it is prudent to avoid "lock-in" of research to a single provider, due to incompatibilities between services provided by different companies. Potentially, a publicly-funded cloud infrastructure would be valuable for running such evaluation experiments in a neutral way.

The proposed infrastructure could also allow evaluation to be conducted on private or restricted data, such as electronic health records or private e-mails, as it is not necessary for the participants to see the test data. However, for researchers, such an approach could be considered unsatisfactory, as researchers would simply obtain metrics on the performance of their algorithms on the data, but not have the possibility to examine why their algorithms performed as they did. Innovative approaches to allow researchers to explore key parts of the private data related to their algorithm performance without revealing private details remain to be developed.

VISCERAL is a solid step toward creating a framework for evaluation of algorithms on large datasets. The framework can be seen as an initial low-level building block of the Innovation Accelerator, as envisioned by van Harmelen et al. [11] as an outline for revolutionising the scientific process.

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements 318068 (VISCERAL), Khresmoi (257528) and 258191 (PROMISE).

References

1. Alonso, O., Baeza-Yates, R.: Design and implementation of relevance assessments using crowdsourcing. In: Advances in Information Retrieval (ECIR 2011), LNCS, vol. 6611, pp. 153–164. Springer (2011)

[18] http://aws.amazon.com/grants/

2. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. Commun. ACM 53(4), 50–58 (2010)
3. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: EvaluatIR: an online tool for evaluating and comparing ir systems. In: SIGIR'09: Proceedings of the 32nd international ACM SIGIR conference. p. 833. ACM (2009)
4. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements that don't add up: ad-hoc retrieval results since 1998. In: CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management. pp. 601–610. ACM (2009)
5. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the Grid: Enabling scalable virtual organizations. The International Journal of Supercomputer Applications 15(3) (Summer 2001)
6. Freire, J., Silva, C.T.: Making computations and publications reproducible with VisTrails. Computing in Science & Engineering 14(4), 18 –25 (Aug 2012)
7. Gagliardi, F., Jones, B., François, G., Bégin, M.E., Heikkurinen, M.: Building an infrastructure for scientific grid computing: status and goals of the EGEE project. Philosophical Transactions of the Royal Society A 363, 1729–1742 (2005)
8. Hanbury, A., Müller, H., Langs, G., Weber, M., Menze, B., Fernandez, T.: Bringing the algorithms to the data: Cloud-based benchmarking for medical image analysis. In: Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics, LNCS, vol. 7488, pp. 24–29. Springer (2012)
9. Hand, D.J.: Classifier technology and the illusion of progress. Statistical Science 21(1), 1–14 (2006)
10. Harman, D.: Information Retrieval Evaluation. Morgan & Claypool Publishers (2011)
11. van Harmelen, F., Kampis, G., Börner, K., Besselaar, P., Schultes, E., Goble, C., Groth, P., Mons, B., Anderson, S., Decker, S., Hayes, C., Buecheler, T., Helbing, D.: Theoretical and technological building blocks for an innovation accelerator. The European Physical Journal Special Topics 214(1), 183–214 (2012)
12. Ince, D.C., Hatton, L., Graham-Cumming, J.: The case for open computer programs. Nature 482(7386), 485–488 (Feb 2012)
13. Langs, G., Müller, H., Menze, B.H., Hanbury, A.: Visceral: towards large data in medical imaging - challenges and directions. In: Proc. MICCAI 2012 Workshop on Medical Content-based Retrieval for Clinical Decision Support (MCBR-CDS) (2012)
14. 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)
15. Pitkanen, M., Zhou, X., Tuisku, M., Niemi, T., Ryynänen, V., Müller, H.: How Grids are perceived in healthcare and the public service sector. In: Global HealthGrid: e-Science Meets Biomedical Informatics — Proceedings of HealthGrid 2008. Studies in Health Technology and Informatics, vol. 138, pp. 61–69. IOS Press (2008)
16. Rebholz-Schumann, D., Yepes, A.J.J., van Mulligen, E.M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC silver standard corpus. Journal of Bioinformatics and Computational Biology 08(01), 163–179 (2010)
17. Sanderson, M.: Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval 4(4), 247–375 (2010)
18. Stodden, V.: The legal framework for reproducible scientific research: Licensing and copyright. Computing in Science & Engineering 11(1), 35 –40 (Feb 2009)

19. Thornley, C.V., Johnson, A.C., Smeaton, A.F., Lee, H.: The scholarly impact of trecvid (2003–2009). Journal of the American Society for Information Science and Technology 62, 613–627 (2011)
20. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of imageclef. In: Proc. CLEF Conference (2011)
21. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1449–1456 (2011)