A Cloud-based Framework for Evaluation on Big Data

Presentation Proposal

Allan Hanbury*, Georg Langs†, Bjoern Menze‡, Henning Müller+

* Vienna University of Technology, Austria

† Medical University of Vienna, Austria

‡ Technische Universität München, Germany

+ University of Applied Sciences Western Switzerland, Switzerland

There are many algorithms, methods and techniques available for Big Data Analytics. This means that there are generally multiple ways to solve a problem, and it is usually not immediately clear from the beginning which combination of algorithms and techniques will produce the optimal solution. The approaches to solving data analytics problems have to be built on a deep understanding of the techniques and methods and how they perform in different situations. Part of this understanding can be gained through extensive evaluation of the techniques and methods.

Currently, whether the optimal solution is found can often depend on the aptitudes and skills of the people solving the problem. Kaggle[1] takes advantage of the dependence on aptitude and skills by crowdsourcing solutions to data analytics problems. Companies post data and associated problems as competitions on the Kaggle platform, and anybody can submit solutions to the problems. Incentives include prize money, fame (people winning multiple competitions are promoted on the Kaggle main page), and occasionally employment offers. However, while such platforms may be a useful source of well-functioning solutions, they do not provide deep understanding of how and why the solutions work and why certain solutions are better than others.

Solutions published in the scientific literature are usually evaluated by comparing a proposed solution to one or more other solutions on one or more datasets selected by the authors. Problematic aspects of this type of evaluation include the use of unsuitable or proprietary datasets, or comparison to poor baselines, leading to “Improvements that don't add up”[2] or an “illusion of progress”.[3] One of the main difficulties is evaluating approaches from multiple research groups on commonly available but realistically large datasets. This is because it is often not straightforward to distribute datasets of several terabytes in size – the usual current approach is to send the data on hard drives by post.

In this presentation, we present a cloud-based framework for evaluation on big data developed in the VISCERAL project.[4] It was developed to allow evaluation of medical imaging algorithms on large amounts of 3D medical imaging data –Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) – and has already been used to run a benchmark on the evaluation of automated segmentation of organs in (fully anonymised) MRI and CT images. An aspect of the framework that makes it attractive for evaluations on medical data is that the data are stored centrally, and it is not necessary to distribute

[1] http://kaggle.com

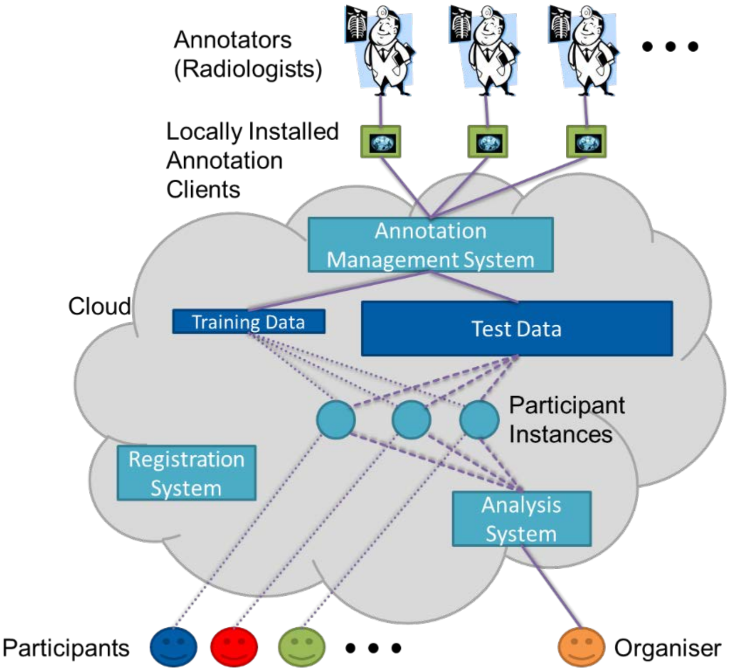
[2] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. *Improvements that don’t add up: ad-hoc retrieval results since 1998*. In CIKM '09: pages 601–610. ACM, 2009.

[3] D. J. Hand. *Classifier technology and the illusion of progress*. Statistical Science, 21(1):1–14, 2006.

[4] http://visceral.eu

the data to multiple locations, where such distribution can be considered as unacceptable from an ethical viewpoint.

The framework, currently running on the Microsoft Azure Cloud, coordinates both the submission of algorithms to evaluate by participants and the manual annotation of images to create ground truth against which the algorithm results are compared. It is shown in the following diagram:



The training data is placed on the cloud. During the training phase of the benchmark (dotted lines above), participants register using the Registration System, and get assigned a Virtual Machine (VM) in the cloud with access to the training data. By the end of the training phase, participants must have installed software completing the provided segmentation task in their VM. During the testing phase (dashed lines above), the organisers of the evaluation take over the VMs and, using the Analysis System, run the installed software on the test data in order to evaluate how well the participant programs perform.

The Annotation Management System manages the manual annotation of the images to create the ground truth. For the medical imaging case, the annotation is done (and can only be done) by qualified radiologists, so the cost of manual annotation is high. The framework assigns the manual annotation resources (radiologists in this case) under the assumption that it is possible to only annotate part of the data with the available annotation resources. Therefore the framework aims at using these resources optimally, by ranking the images and organs to identify those for which annotation would mean the maximum information gain, while at the same time performing quality checks of the annotations. The framework generates annotation “tickets” for specific annotators that specify the image and anatomical structures to be annotated, and accepts the upload of finished annotations.

While this framework has been developed for evaluation on medical imaging data, it is general enough to be used on other types of data. Work is currently underway to automate as much of the evaluation process as possible, to reduce the interventions required by the evaluation organisers to a minimum, and hence to provide more rapid feedback on evaluation outcomes to the participants.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 318068 (VISCERAL).