Final Assignment

"Artificial Scientist and the tools needed to get there"

Science Theory and Research Methodology DD2205

Jim Holmström 890503-7571
jimho@kth.se

November 24, 2012

1 Summary

This essay will be focusing on the current software and hardware limitations towards the development of new machine learning techniques such as the artificial scientist. Much of the claims and discussions will be based on the article "Towards 2020 Science" by Microsoft Research.[1]

It seems that we are currently standing even closer to the brink of a new computer revolution. New techniques are starting to be developed and set into production to overcome new types of hinders of computing power. This hinder is is nowadays much more focused on how to actual distribute the computations since the actual smallest computing unit is starting to hit the so called heat wall.

2 The dawn of new hardware

Super scalability * Hardware isn't really catching up with the datasets, clock speed and memory bandwidth has stagnated and the only way to increase the computing power is to increase the number of cores or change the underlying architecture (as in GPU) but this introduces a lot of new problems one has to re-implement all algorithms for the new setup efficiently which is non trivial in many cases. Mention cache-locks and

* Hadoop (with Mahong ofc) is the most widely used framework for distributed processing on immense datasets across a cluster.

* Non-perfect hardware: many calculations need not to be perfect to be able to operate and some research into this has been done: //REF

3 A new breed of scientists

Artificial Scientist: * Use algorithms with active learning on real life experiments for example the microfluid one. * Use an artificial scientist to improve the artificial scientist, OMFG

* Hardware is a bit behind, need neurla processors or processors which do errors but the algorithms are errorcorrecting initself, which is closer to neural-processing in the neuronnets. * also need other ways of thinking dealing with massive-parallel, refer to bigdata-eventguy which had had the new parallel-databse idea. Most machinelearning algorithms is non-sequencial(is it called this?) but separating the problem in a map-reduce fasion isn't always trivial. * An important piece of the puzzle is to get machinelearning to be scaleable in deeplearning(is this correct usage of deeplearning?) so that we don't have to generate features byhand but instead have raw input with minimal preprocessing (perhaps only log-scaling some raw input you know have a exponential behavour.

* Or redesigning the hardware it is running on to be optimal to .. (recursive call)

References

[1] Microsoft Research. Towards 2020 Science. `research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf`.