A propos du centre ou de la direction fonctionnelle

The Inria Sophia Antipolis – Méditerranée center counts 37 research teams and 9 support departments. The center’s staff (about 500 people including 400 Inria employees) is composed of scientists of different nationalities (250 Foreigners of 50 nationalities), engineers, technicians and administrators. 1/3 of the staff are civil servants, the others are contractual. The majority of the research teams at the center are located in Sophia Antipolis and Nice in the Alpes-Maritimes. Six teams are based in Montpellier and a team is hosted by the computer science department of the University of Bologne in Italy. The Center is a member of the University and Institution Community (Comité) “Université Côte d’Azur (UCA)”.  

Contexte et atouts du poste

The post-doc will take place in the NEO project-team [https://team.inria.fr/neol]. NEO is positioned at the intersection of Operations Research and Network Science. By using the tools of Stochastic Operations Research, the team members model situations arising in several application domains, involving networking in one way or the other. The aim is to understand the role of the effects in order to influence and control them so as to engineer the creation and the evolution of complex networks.  

The research activity will be supervised by Giovanni Neglia [http://www-sop.inria.fr/members/Giovanni.Neglia/].

Mission confiée

CONTEXT, POSITIONING AND OBJECTIVES

The recent success of machine learning techniques such as deep learning stems from a number of fundamental algorithmic improvements [11, 15]. Since most of the fundamental techniques underlying modern machine learning have their roots in the 80’s, what really makes the difference, nowadays, is the availability of unprecedented amounts of training data which, coupled with the commoditization of general purpose GPUs, makes it possible to build models capable of achieving accuracies that in some cases surpass human ability [20]. While a single computer equipped with powerful GPUs provides algorithmic simplicity and speed up to a given scale of training data and model size [11, 10, 3], today we have reached an operating point where both datasets and model sizes no longer fit the capacity of a single machine. Thus, a distributed implementation of training algorithms for modern statistical models, such as deep learning, becomes truly necessary.

In general, the success of deep learning as well as modern Bayesian non parametrics [8], stems from the application of stochastic optimization algorithms [2, 8, 14], that iteratively work on randomized subsets of the training data to determine model parameters that minimize a given loss function. Many learning problems can be expressed as (stochastic) optimization problems: given a dataset X = {x1, x2, ... xn}, the goal is to learn the parameters θ of a model with respect to an empirical loss function L(θ, X).

A first-order stochastic optimization algorithm achieves this by iteratively updating using a stochastic gradient of L(θ, X) computed at a randomly sampled xi. This iterative procedure, called Stochastic Gradient Descent (SGD), produces a sequence of models, that is guaranteed to converge to a local optimum of the loss function [18]. Stochastic optimization can be seen as a unifying methodology to attack general learning problems at an unprecedented scale. It is therefore important to define novel algorithms and synthesize improvement to distribute optimization algorithms across a cluster of machines that cooperate to minimize a given loss function. Despite several efforts in this direction [4, 16, 1], distributed optimization is still in its infancy, and more research is required to improve performance, scalability, and generality of optimization algorithms.

REFERENCES

GOALS

Apache Spark represents a natural candidate for large scale statistical modeling, due to its natural predisposition to process distributed data structures (which can hold model parameters as well as gradient updates). Nevertheless, Apache Spark is built on the Bulk Synchronous Parallel processing model [21], which currently restricts its application to synchronous optimization algorithms.

The performance of Spark can significantly be affected by “stragglers.” Only recently the problem of stragglers has been addressed, using a deceivingly simple approach. Regular workers can be complemented by “backup” workers, which contribute to the computation of gradient updates [3]. More formally, given a set of $n$ workers and $b$ backup workers, the parameter server coordinate a total of $n + b$ workers, but waits for any $n$ gradient updates for computing the next gradient aggregate. Although in practice this approach has shown to partially mitigate the problems of slow workers, it suffers from three main problems. First, depending on the value of $b$, a lot of redundant work is eventually lost. Second, depending on low-level infrastructure details, backup workers might be scheduled concurrently to regular workers on the same machine, which artificially creates resource contention, thus producing new “stragglers.” Third, determining an appropriate number of backup workers for a given setup (which includes also the particular statistical model being optimized) remains elusive. In this research project we aim to address such issues, by working on efficient speculative execution mechanisms that avoid wasting work, while making sure to mitigate the “straggler” problem rather than possibly exacerbating it.

Principales activités

Research activity.

Possible supervision of interns.

Compétences

Competences in probability, statistics, optimization, and mathematical modeling are essential. Solid programming and IT skills are necessary, along with strong communication abilities.

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Rémunération

Gross Salary: 2650 brutto per month