2018-00528 - [CORDIS2018-ZENITH] Learning models parameters at scale

Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle
Le centre Inria Sophia Antipolis - Méditerranée compte 37 équipes de recherche, ainsi que 9 services d’appui à la recherche. Le personnel du centre (600 personnes environ dont 400 salariés Inria) est composé de scientifiques de différentes nationalités (250 personnes étrangères sur 50 nationalités), d’ingénieurs, de techniciens et d’administratifs. 1/3 du personnel est fonctionnaire, les autres sont contractuels. La majorité des équipes de recherche du centre sont localisées à Sophia Antipolis et Nice dans les Alpes-Maritimes. Six équipes sont implantées à Montpellier et une équipe est hébergée par le département d’informatique de l’université de Bologne en Italie. Le Centre est membre de la Communauté d’Université et d’Établissement (ComUE) « Université Côte d’Azur (UCA) ».

Contexte et atouts du poste

The team and advisors
The Inria Zenith team is dedicated to large-scale scientific data management. The PhD will be supervised by Patrick Valduriez, a senior researcher at Inria working on distributed and parallel data management [9], and by Antoine Liutkus, a researcher at Inria working on machine learning and music signal processing [10, 11].

Insertion in the team
The defining objective of this PhD will be scaling up the estimation of probabilistic models thanks to the sketching methodology, applied to the management of massive scientific data, reflecting the team's specificities.

While music and audio heritage preservation is envisioned as the main application, the successful candidate will also apply his findings to the other applicative domains considered by the team, such as natural heritage preservation and indexing [12].

Collaborations
The PhD student will furthermore benefit from a dense network of scientific collaborators. In particular:
- The CREM (Centre de Recherches en Ethnomusicologie) is the UMR responsible for the Archives du CNRS Musée de l’Homme, that will be the main dataset considered for this Ph.D. It is a leading laboratory for the preservation of our immaterial audio heritage.
- The PANAMA team, lead by Rémi Gribonval, is a leading place for research on sketching approaches. The recruited Ph.D student will have exchange stays in this team.
- Telecom Paristech is a leading place for data science. As a renowned researcher for models parameters estimation with MCMC, Umut Şimşekli is envisioned as a privileged partner. As an internationally renowned specialist for audio processing, Roland Badeau is also envisioned as a close collaborator.

Mission confiée

Context
The activity of researchers often revolves around analyzing data to learn model parameters that are used by applications. However, the availability of big data requires a shift in the tools used for this purpose.
Deep learning [1] is emerging as an appropriate methodology to explore, analyze, and leverage big data. While initially framed as a sequential procedure, much research on deep learning is currently focusing on parallel and distributed architectures [2, 3] that allow scaling up to match the needs of scientists. This is first achieved by enabling asynchronous optimization strategies, which alleviate the sequential nature of training to some extent [4, 2], and also by distributing either the data or parts of the model to multiple nodes. In any case, these strategies clearly exhibit some shortcomings: the computing gain is rapidly mitigated by the communication cost between nodes. Parallelizing over more than a dozen nodes is still a challenge [5].

Assignments

This PhD project will investigate new ways of training model parameters in a distributed and parallel fashion, by exploiting the recently proposed sketching methodology [6, 7]. Its core idea is to drop the need to analyze the actual (massive) data, but rather focus on summary statistics computed beforehand. Recent research showed that effective sketching strategies allow for the provably correct estimation of model parameters in some cases and applied the method to various machine learning and signal processing tasks [8, 7].

Here, we will investigate the impact of such a strategy on the large-scale learning of large data models, including but not limited to deep neural networks. The fundamental fact to be exploited is that sketching and learning can be performed in a parallel and totally asynchronous way.

Applications

This PhD will use large-scale learning for the automatic restoration and browsing of audio ethnomusicological archives. This topic is at the crossroads of cultural heritage preservation, large-scale learning, audio and music signal processing and probability theory.

Through previous fundings, the Zenith team is developing cutting-edge audio restoration techniques and making them available for research purpose in an open-source environment embedded in real-world historical sound archives. The Ph.D candidate will go further into bridging the gap between research in computer science and digital humanities, striving for the preservation of our immaterial heritage. The intended outcome are audio analysis and processing tools designed through systematic training on the 50k items of the CNRS-Musée de l'Homme archives.

The successful candidate will conduct theoretical and applied research, with an expected impact on both the machine learning and the digital humanities communities.

References


Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.
Principales activités
The tasks to be realized by the PhD student will be to:

1. Investigate the state-of-the-art approaches in parallel deep learning
2. Propose novel, efficient parallel techniques based on the sketching methodology
3. Validate the techniques by building a prototype on a parallel platform and performing experiments on big datasets

Compétences
Additional, useful elements are:

- Programming experience with Hadoop or any software framework for distributed computing like Spark and Scala.
- Programming experience with Python and using GPU.
- Notions of signal processing and machine learning.
- A strong interest in music and cultural heritage preservation

Avantages sociaux
- Restauration subventionnée
- Transports publics remboursés partiellement
- Sécurité sociale
- Congés payés
- Aménagement du temps de travail

Rémunération
Durée: 36 mois
Localisation: Sophia Antipolis, France
Rémunération: 1982€ brut mensuel (année 1 & 2) et 2085€ brut mensuel (année 3)