About the research centre or Inria department
The Inria Sophia Antipolis - Méditerranée center counts 36 research teams as well as 8 support departments. The center's staff (about 500 people including 320 Inria employees) is made up of scientists of different nationalities (250 foreigners of 50 nationalities), engineers, technicians and administrative staff. 1/3 of the staff are civil servants, the others are contractual agents. The majority of the research teams that are located in Sophia Antipolis and Nice in the Alpes-Maritimes. Four teams are based in Montpellier and two teams are hosted in Bologna in Italy and Athens. The Center is a founding member of Université Côte d'Azur and partner of the i-Site MUSE supported by the University of Montpellier.

Project description
MAMMALS aims to provide low-latency inferences by running—close to the end user—simple machine learning models that can also take advantage of a (small) local datastore of examples. The focus is on algorithms to learn online what to store locally to improve inference quality and achieve domain adaptation.

MAMMALS provides a framework to design algorithms to learn online what to store locally to improve inference quality and achieve domain adaptation. It assumes that the data is available locally at the edge and that the model can be run locally.

Assignment
The post-doc will work with a PhD student already hired.

Instruction to apply
Defence Security: This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy: As part of its diversity policy, all Inria positions are accessible to people with disabilities.

Warning: You must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

2021-03582 - Post-Doctoral Research Visit F/M Memory-augmented Models for low-latency Machine-learning Serving

Contract type: Fixed-term contract
Level of qualifications required: PhD or equivalent
Function: Post-Doctoral Research Visit

REFERENCES
Main activities

In the framework of the project described above, the postdoc can work on a combination of the three following aspects.

I. Design of online learning algorithms.

We plan to evaluate three different frameworks for learning online how to populate the local datastore.

1. Adapt existing caching policies like LRU, e.g., by inserting the content on the basis of its marginal utility (i.e., its contribution to inference quality). Ad-hoc policies in this spirit have been proposed to support image similarity search in [5] and in [8]. This framework leads usually to a combinatorial analysis with a focus on expected performance under a stochastic request process.

2. Study the problem as a discrete-space metrical task system (MTS) [2], where the state of the system is the set of instances in the datastore. Each state has a corresponding service cost (the loss of inference quality due to running a simpler model at the edge) and updating the datastore generates so-called movement costs. Competitive analysis is the common approach to study this setting.

3. When the set of possible instances is very large and roughly homogeneously distributed, at least over a low-dimension manifold, it is possible to consider the state space to be continuous. This setting is closer to online machine learning with regret as its main performance metric.

At the methodological level, we will explore gradient-based approaches. They are common in online machine learning, but, more recently, they have also been effectively employed to study combinatorial problems in the other two settings [3, 3, 14].

II. Characterization of datasets' topological properties.

Which framework, among the three described above, is the most appropriate? The answer depends on a large extent on the topological properties of the space where instances lie. Whereas we are looking for collaborations with other research teams studying the topological and geometric structure of data, we will push a practical approach, starting from real traces. Many traces are available for recommender systems based on ML predictors. This application is particularly interesting for MAMMALS, as recommendations need to be customized to the user (a particular example of domain adaptation) and constantly updated to follow dynamic popularities of media contents or products.

III. Prototype implementation.

We plan to provide practical evidence of the potential improvements from MAMMALS new algorithms in a simpler context. In many ML and information retrieval applications it is required to retrieve fact the k nearest neighbours (k-NN) of a given point in a dataset. When the number of dimensions exceeds 10, exact k-NN computation essentially requires to scan the whole dataset [17]. To simplify this problem, we will study the problem as a discrete-space metrical task system (MTS) [2], where the state of the system is the set of instances in the datastore. Each state has a corresponding service cost (the loss of inference quality due to running a simpler model at the edge) and updating the datastore generates so-called movement costs. Competitive analysis is the common approach to study this setting.

When the set of possible instances is very large and roughly homogeneously distributed, at least over a low-dimension manifold, it is possible to consider the state space to be continuous. This setting is closer to online machine learning with regret as its main performance metric.

At the methodological level, we will explore gradient-based approaches. They are common in online machine learning, but, more recently, they have also been effectively employed to study combinatorial problems in the other two settings [3, 3, 14].

Skills

We are looking for one of the following profiles:

1) a candidate with solid analytical skills to design algorithms with strong performance guarantees,
2) a candidate expert on high-dimensional data analysis,
3) a candidate with hands-on experience on machine learning, able to reproduce state-of-the-art results like those in [12] and in [19].

Benefits package

- 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

Remuneration

Gross Salary: 2633 € per month