Offer #2020-02866

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

Fonction : Post-Doctoral Research Visit

About the research centre or Inria department

The Inria Sophia Antipolis - Méditerranée center counts 34 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 center's research teams 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.

Context

The post-doc will take place in the NEO project-teamhttps://team.inria.fr/neo/.

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

The research is in the framework of the Inria's exploratory action MAMMALS (Memory-augmented Models
for low-latency Machine-learning Serving). The postdoc will collaborate with a PhD student already hired.

Assignment

The research is in the framework of the Inria's exploratory action MAMMALS (Memory-augmented Models
for low-latency Machine-learning Serving) described below.

SUMMARY
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.

PROJECT DESCRIPTION
A machine learning (ML) model is often trained for inference's purposes. Inference does not involve complex iterative algorithms and is therefore generally presumed to be easy. Nevertheless, it presents fundamental challenges that are likely to become dominant as ML adoption increases and ML systems are ubiquitously deployed and need to make timely and safe decisions in unpredictable environments [16]. Big cloud providers, such as Amazon, Microsoft, and Google, offer their "machine learning as a service" (MLaaS) solutions, but running the models in the cloud may fail to meet delay constraints. As an example, recommendation systems, voice assistants, and ad-targeting need to serve predictions in less than 20 ms. Future 5G wireless services for connected and autonomous cars, industrial robotics, mobile gaming, augmented and virtual reality have even stricter latency requirements, often below 10ms and below 1 ms for what is now called the tactile internet [15]. Such requirements can only be met by running ML prediction services at the edge of the network—directly on users' devices or at nearby servers—without the computing and storage capabilities of the cloud. Privacy and data ownership also call for inference at the edge.
The current approach to run inference at the edge is to take state-of-the-art (SOTA) large ML models (often neural networks) and generate smaller ones through compression or distillation [6]. MAMMALS will pursue a different di- rection. Its key idea is to take advantage of data availability at the edge (where data is usually generated) to compensate for additional computing constraints. In particular, we want to combine the decisions of a small ML model, e.g., a compressed neural network, with those of an instance-based algorithm relying on a local datastore, like k-nearest neighbors (k-NN). Instance-based algorithms can explicitly memorize rare patterns that are difficult to learn by simple ML models. Moreover, they do

not require complex training and can efficiently incorporate new information.
This activity builds on some recent findings showing that ML models can benefit from the presence of a local datastore or memory. Inspired by the (complex) memory-augmented neural networks [10, 11], some recent papers [9, 13, 12], have shown that the performance of SOTA neural networks can benefit from a memory storing a simple collection of examples, from which the most similar ones to the current input are retrieved to improve neural network inferences. These results are quite surprising as “in the machine learning research community it is generally believed that there is a tension between memorization and generalization” [4]. MAMMALS will exploit this synergy of model-based and instance-based learning to achieve more flexibility in adapting inference engines to limited edge resources.
Instances selection is a challenging task, and MAMMALS indeed focuses on designing online algorithms to decide what to store locally. Note that this corresponds to train the instance-based algorithm.

REFERENCES
[1] J. Abernethy et al. “A Regularization Approach to Metrical Task Sys- tems”. In: Algorithmic Learning Theory. 2010.
[2] A. Borodin et al. “An Optimal On-line Algorithm for Metrical Task Sys- tem”. In: J. ACM 39.4 (Oct. 1992).
[3] S. Bubeck et al. “K-Server via Multiscale Entropic Regularization”. In: ACM STOC. 2018.
[4] S. Chatterjee. “Learning and Memorization”. In: Proceedings of the 35th International Conference on Machine Learning. Vol. 80. Oct. 2018.
[5] F. Chierichetti et al. “Similarity Caching”. In: ACM PODS. 2009.
[6] B. L. Deng et al. “Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey”. In: Proc. of the IEEE (2020).
[7] Faiss: A library for efficient similarity search. https : / / github . com / facebookresearch/faiss/.
[8] M. Garetto et al. “Similarity Caching: Theory and Algorithms”. In: IEEE INFOCOM. 2020.
[9] E. Grave et al. “Unbounded Cache Model for Online Language Modeling with Open Vocabulary”. In: NIPS. 2017.
[10] E. Grefenstette et al. “Learning to Transduce with Unbounded Memory”. In: NIPS. 2015.
[11] A. Joulin and T. Mikolov. “Inferring Algorithmic Patterns with Stack-augmented Recurrent Nets”. In: NIPS. 2015.
[12] U. Khandelwal et al. “Generalization through Memorization: Nearest Neighbor Language Models”. In: ICLR. 2020.
[13] S. Merity et al. “Pointer Sentinel Mixture Models”. In: ICLR. 2017.
[14] G. S. Paschos et al. “Learning to Cache With No Regrets”. In: IEEE INFOCOM. 2019.
[15] M. Simsek et al. “5G-Enabled Tactile Internet”. In: IEEE Journal on Selected Areas in Communications 34.3 (2016).
[16] I. Stoica et al. A Berkeley View of Systems Challenges for AI. Tech. rep. UCB/EECS-2017-159. Oct. 2017.
[17] R. Weber et al. “A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces”. In: VLDB. 1998.
[18] D. R. Wilson and T. R. Martinez. “Reduction Techniques for Instance- Based Learning Algorithms”. In: Mach. Learn. 38.3 (Mar. 2000).
[19] J. Johnson, M. Douze, M., H. Jégou, “Billion-scale similarity search with GPUs”. IEEE Transactions on Big Data (2019)

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 [1, 3, 14].

II. Characterization of datasets' topological properties.

Which framework, among the three described above, is the most appropriate? The answer depends to 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 fast 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], so specialized approximate indexing structures have been proposed and are currently implemented in libraries like Facebook Faiss [7]. Now, these systems can also benefit from a fast memory that stores a small subset of the whole repository. Managing this memory dynamically presents many of the challenges described above with the advantage of 1) avoiding the additional complexity of the interaction with the model, and 2) having a clear evaluation framework with well established benchmarks and performance metrics.

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: 2653 € per month

General Information

- **Theme/Domain :** Networks and Telecommunications
  System & Networks (BAP E)
- **Town/city :** Sophia Antipolis
- **Inria Center :** Centre Inria d'Université Côte d'Azur
- **Starting date :** 2020-10-01
- **Duration of contract :** 1 year, 6 months
- **Deadline to apply :** 2020-11-30

Contacts

- **Inria Team :** NEO
- **Recruiter :**
  Neglia Giovanni / Giovanni.Neglia@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

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.

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.