2019-01436 - Post-Doctoral Research Visit F/M Low-Latency Machine Learning Prediction Services

About the research centre or Inria department
Inria is a national research institute dedicated to digital sciences that promotes scientific excellence and transfer. Inria employs 2,400 collaborators organised in research project teams, usually in collaboration with its academic partners. This agility allows its scientists, from the best universities in the world, to meet the challenges of computer science and mathematics, either through multidisciplinarity or with industrial partners. A precursor to the creation of Deep Tech companies, Inria has also supported the creation of more than 150 start-ups from its research teams. Inria effectively faces the challenges of the digital transformation of science, society and the economy.

Context
This postdoc is in the framework of Nokia Bell Labs - Inria joint lab. The research activity will be carried out at Inria Sophia Antipolis Méditerranée, but periodic visit to Nokia Bell Labs, Paris, are envisaged.

How to apply:
Applications will be evaluated on a rolling basis. Starting date could be as early as May 2019. Send to Giovanni Neglia (giovanni.neglia@inria.fr) your application containing:
- CV
- Motivation letter
- PhD thesis if already completed,
- Four most representative publication
- Contact information of potential referees

Assignment
Many deployed applications, like recommendation systems, voice assistants, and ad-targeting, need to serve predictions from machine learning (ML) models in less than 20ms [1]. Future wireless services like connected and autonomous cars, industrial robotics, mobile gaming, augmented and virtual reality have even stricter latency requirements, often below 10 ms [2] and below 1ms for what is now called the tactile internet [3]. A key element to satisfy such constraints is to run these services closer to the user.

In particular, it will be needed to run ML prediction services at the edge of the network without the computing and storage capabilities of the cloud. The postdoc will investigate how the quality of the predictions can be traded off with latency through two different approaches. The first one is to cache at the edge ML answers to previous queries. Cached answers to “close enough” queries can then be provided to new queries. Local sensitive hashing is a possible way to evaluate the distance between queries [4]. The second approach is instead to train a complex ML model in the cloud, but then run a down-scaled version at the edge. To this purpose, transfer learning techniques can be used to generate models with different time-space complexity [5].

The duration of this postdoc is 18 months.


Main activities
Research activities.

Skills
Competences in probability, statistics, optimization, and mathematical modeling are essential. Solid programming and IT skills are necessary, along with strong communication abilities.
Ideally, we are looking for two possible profiles:
- experts on algorithms and competitive analysis
- experts on machine learning

Benefits package
- Subsidized meals
- Partial reimbursement of public transport costs
- Leaves 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: 2650 brutto per month