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 1ms 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 deep networks) and generate smaller ones through compression or distillation [6]. MAMMALS will pursue a different direction. 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 new information.

A propos du centre ou de la direction fonctionnelle

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.

Contexte et atouts du poste

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

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

Mission confiée

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

Informations générales

- Thème/Domaine : Réseaux et télécommunications
- Systèmes & réseaux (BAP E)
- Ville : Sophia Antipolis
- Centre Inria : CRI Sophia Antipolis - Méditerranée
- Date de prise de fonction souhaitée : 2020-10-01
- Durée de contrat : 1 an, 6 mois
- Date limite pour postuler : 2020-11-30

Contacts

- Equipe Inria : NEO
- Recruteur : Neglia Giovanni / Giovanni.Neglia@inria.fr

A propos d’Inria

Inria est l’Institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 200 équipes-projets agiles, en général communiques avec des partenaires académiques, impliquent plus de 3500 scientifiques pour relever les défis du numérique, souvent à l’interface d’autres disciplines. L’institut fait appel à de nombreux talents dans plus d’une quarantaine de métiers différents. 900 personnels d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde.

Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 180 start-up. L’institut s’efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l’économie.

Consignes pour postuler

Sécurité défense :

Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

Politique de recrutement :

Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.

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.
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 benchmarks and performance metrics.

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 small subset of the whole repository. Managing this memory dynamically presents many of the combinatorial problems in the other two settings [1, 3, 14].

At the methodological level, we will explore gradient-based approaches. They are common in online adaptation) and constantly updated to follow dynamic popularities of media contents or products.

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.

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 combinatorial problems in the other two settings [1, 3, 14].

At the methodological level, we will explore gradient-based approaches. They are common in online adaptation) and constantly updated to follow dynamic popularities of media contents or products.

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.

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 combinatorial problems in the other two settings [1, 3, 14].

At the methodological level, we will explore gradient-based approaches. They are common in online adaptation) and constantly updated to follow dynamic popularities of media contents or products.

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.
- 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: 2653 € per month