Offre n°2024-07630

PhD Position F/M (campagne) Embedded AI: Graph Neural Networks in Lossy Embedded Wireless Networks

Type de contrat : Fixed-term contract

Niveau de diplôme exigé : Graduate degree or equivalent

Fonction : PhD Position

Contexte et atouts du poste

Neural Networks make advanced predictions on data, such as identifying objects on a picture. The ones using graph data as input are called Graph Neural Networks (GNNs). As Neural Networks are computationally expensive, these models are usually run on powerful servers. The availability of hardware acceleration modules as well as recent compression techniques suggest it may be possible to run GNNs on low-power wireless networks of constrained embedded devices. This would open up many new applications for GNNs, includin advanced distributed sensing on autonomous swarms of mobile robots. The goal of this PhD is to explore that opportunity. One challenge is that the connectivity between such embedded devices is lossy: messages are lost because of physical phenomena such as external interference and multi-path fading. A second challenge is that, in some use cases such as swarm robotics, real-time constraints come into play. A third challenge is how to retrain and update the model in a setting where transferring large amounts of data to a central server is prohibitive in terms of time and energy. This research is exhilarating and high-risk high-gain by nature; it has the potential of redefining what embedded AI means, and open tremendous opportunities for applications in distributed systems, robotics and robot swarms.

Supervised learning is one of the main branches in Artificial Intelligence (AI). Training a model requires an ensemble of data to train on. Each data item is associated with a label, which is the best prediction possible. Model produces one output per data item in a suite of operations called inference; the differences between the label and the inference result are used to adjust the model parameters, until its results are sufficiently accurate. It can then be used to make predictions, for example detecting spoken words in the audio samples.

Neural Networks is a category of supervised learning models. Its main interest lies in its capacity to make complex predictions at a level of accuracy superior to any other prediction algorithm, in a great variety of tasks. Typically, data is collected by an embedded system and sent to a remote server, where the AI model is trained. The model is then compressed [1] and transferred for inference purposes in the embedded systems. This way the devices can make predictions locally over the data they just collected; this is called on-edge inference. Recent progress makes it even possible to fine-tune the model by training it on device, to improve its accuracy and adapt to a changing context.

When a Neural Network specifically uses graph data as its input, it's called a Graph Neural Network (GNN). It was created because it can train and make inference over graphs of different sizes, while other neural networks rely on fixed-size inputs. That property makes it suitable for communication networks among others, where the number of devices and their connections varies over time.

GNNs are usually running on a server in the cloud, where computational resources are abundant. Distributing a GNN to a network of resource-limited embedded devices would tremendously increase the devices intelligence.

Inference in GNNs consists of two steps. In the first step, each node exchanges messages with its surrounding nodes and computes a fixed-size output using an aggregation function (e.g. mean). In the second step, each device turns the output of the aggregation function into a prediction. Because all devices are sharing information, each device has a bigger part of the graph information than if it was just using its own data, which leads to better predictions.

In such a low-power wireless embedded scenario, the wireless links are unreliable in nature. There is hence a non-zero probability that messages are not received because of the distance between devices, of interference from other devices, or of phenomena such as multi-path fading.

Intuitively, such communication unreliability impacts the quality of the prediction. That being said, the impact of that unreliability is not well studied on GNNs as existing work assumes ideal communication.

Mission confiée

The Grand Challenge of the proposed PhD work is to explore the impact of lossy communications on the performance of Graph Neural Networks in embedded systems networks, where devices are constrained by their available memory, computation speed and battery capacity.

This Grand Challenge translates into three Scientific Objectives.

Scientific Objective 1: finding techniques to mitigate the negative effects lossy networks have on the performance of GNNs. This is capital for real-world efficiency of GNNs on embedded systems, and serves as a foundation for the remainder of the work. What is the right trade-off between accuracy, latency and energy consumption?

Scientific Objective 2. In most embedded systems, predictions and other tasks must be achieved in a limited time, reducing the possible corrections to take the right decision (e.g. waiting more computations or information to confirm). Can the message passing phase be adapted so every device has sufficient information even if some messages are lost? Would combining the device old data with its current data improve its accuracy ?

Scientific Objective 3. Updating the GNN model with data collected after the devices have already been deployed would be an excellent way to improve its adaptability, also preventing performance drop due to environment evolution. The challenge of course is that the capacity of such networks is very limited. Could the model be improved by training it on the devices, to remove dependency to a remote server?

Answering these three scientific objectives opens up tremendous opportunities for real-world practical use cases. A swarm of low-power wireless autonomous robots could operate as one large distributed microphone, implementing a GNN to collectively recognize and classify sounds, raising an alarm when detecting a dangerous situation.

Principales activités

The research outlined in this document is tailored for a 36 month doctoral research program.

Year 1. The main objective of year 1 is to explore the state of the art related to this topic, both in terms of academic literature, practical use cases and implementation, and research community. This work will result in the submission of a survey paper. You will start exploring SO1 at the end of year 1. You will look at how GNN structure can be adapted to mitigate negative effects occurring in lossy environments.

Year 2. Answering the questions of Scientific Objectives 2 and 3 is the goal of year 2. In SO2, you will consider lossy environments coupled with real-time constraints. This will result in at least one conference publication. In SO3, you will look at how mitigating the effects of lossy environments enhances model training on the devices. This will result in at least one conference publication.

Year 3. Your work will culminate into a practical implementation, deployment and real-world test at the very beginning of year 3. You will apply the previously discovered techniques to show the improvements it brings in a swarm of robots. This will result in the submission of an capstone paper to a journal. Finally, the last 6-9 month of year 3 are dedicated to producing you PhD manuscript, submit it to a jury panel, and defending it.

Avantages

- 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 and flexible organization of working hours (after 12 months of employment)
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Rémunération

According to civil service salary scales

Informations générales

- **Thème/Domaine :** Networks and Telecommunications
  System & Networks (BAP E)
- **Ville :** Paris
- **Centre Inria :** Centre Inria de Paris
- **Date de prise de fonction souhaitée :**2024-10-01
- **Durée de contrat :**3 years
- **Date limite pour postuler :**2024-05-19

Contacts

- **Équipe Inria :** AIO

- **Directeur de thèse :**
  Watteyne Thomas / thomas.watteyne@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 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 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 200 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.

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.

Consignes pour postuler

In your application (which can be in English or in French), please include:

- CV
- Letter of motivation
- Letters of recommendation
- Master's grades

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.