Offre n°2025-08930

PhD Position F/M Theoretically justified neural network compression

Type de contrat : Fixed-term contract

Niveau de diplôme exigé : Graduate degree or equivalent

Fonction : PhD Position

Contexte et atouts du poste

The PhD position will be funded by ERC DYNASTY (Umut Simsekli).

The position might include traveling to conferences for paper presentation. Travel expenses will be covered within the limits of the scale in force.

Mission confiée

With the increasing model sizes in deep learning and with its increasing use in low-resource environments, network compression techniques are becoming ever more important. Among many network compression techniques, network pruning has been arguably the most commonly used method [1], and it is rising in popularity and success [2].

A common conclusion in pruning research is that overparametrized networks can be greatly compressed by pruning with little to no cost at the overall performance of the network, including with simple schemes such as magnitude pruning [1,2]. For example, research on *iterative magnitude pruning* [3] demonstrated the possibility of compressing trained deep learning models by iteratively eliciting a much sparser substructure.

Recently, it has been illustrated that the choice of training hyperparameters such as learning rate affects the performance of such pruning strategies [3,4,5,6] took the first step towards providing a theoretical justification for these observations. Yet, it is still highly nontrivial to understand when a pruning method will or will not be

useful [1]. Hence developing compression techniques with strong theoretical guarantees is crucial.

The goal of this thesis will be to address the aforementioned issues. More precisely, it will aim at addressing the following points:

- Make a selective bibliography in the topic and identify most relevant used methods in the existing literature
- Develop new compression techniques by noise injections to optimization algorithms.
- Perform theoretical analysis on the performance of the developed algorithm.
- Support the findings by experiments conducted on several benchmarks in modern neural networks.

References

[1] Neill, James O. "An overview of neural network compression." *arXiv preprint arXiv:2006.03669* (2020).
[2] Blalock, Davis, et al. "What is the state of neural network pruning?." *Proceedings of machine learning and systems* 2 (2020): 129-146.
[3] Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." *arXiv preprint arXiv:1803.03635* (2018).
[4] Zhou, Hattie, et al. "Deconstructing lottery tickets: Zeros, signs, and the supermask." *Advances in neural information processing systems* 32 (2019).
[5] Renda, Alex, Jonathan Frankle, and Michael Carbin. "Comparing rewinding and fine-tuning in neural network pruning." *arXiv preprint arXiv:2003.02389* (2020).
[6] Barsbey, Melih, et al. "Heavy tails in SGD and compressibility of overparametrized neural networks." *Advances in Neural Information Processing Systems* 34 (2021): 29364-29378.

Principales activités

Main activities :

- Conduct theoretical research
- Conduct experiments for empirical verification
- Write scientific articles
- Disseminate the scientific work in appropriate venues.

Compétences

Technical skills and level required :
Languages : High-level of professional/academic English
Coding skills : Good level of coding in Python and related deep learning libraries

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
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Informations générales

- **Thème/Domaine :** Optimization, machine learning and statistical methods Statistics (Big data) (BAP E)
- **Ville :** Paris
- **Centre Inria :** Centre Inria de Paris
- **Date de prise de fonction souhaitée :** 2025-10-01
- **Durée de contrat :** 3 years
- **Date limite pour postuler :** 2025-06-22

Contacts

- **Équipe Inria :** SIERRA
- **Directeur de thèse :**
  Simsekli Umut / umut.simsekli@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'e?orce 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

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.