

Projet d'Initiation à la Recherche

Nom du laboratoire ou de l'entreprise/établissement : LAAS/INSA

TUTEUR(S)

NOM-Prénom

Tel

Mel

KŁODA Tomasz

tomasz.kloda@laas.fr

TITRE DU PROJET

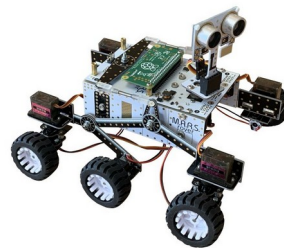
Edge AI for autonomous mini rover

MOT-CLES

Embedded AI, real-time systems, rover

DESCRIPTIF (RESUME)

*Deploying large AI models on edge devices presents several challenges due to the restrained hardware resources and the ever-growing size of AI models. This project aims to develop an **AI perception and control modules for a mini autonomous rover**, focusing on techniques that make AI algorithms feasible in resource-constrained and battery-powered environments. To achieve this, we will explore and integrate the following approaches:*



- **Distributed inference:** Partitioning the model execution across the rover's onboard computer and more computer powerful remote station with hardware accelerators (e.g., part of the model running on the rover while the remainder executes in the remote station) [1].

- **Model compression:** Applying techniques such as quantization (reducing the precision of model parameters) and pruning (removing redundant weights or structures) to obtain a smaller, faster, and more energy-efficient model, while carefully managing the accuracy trade-off [2, 3].

- **Dynamic neural networks:** Implementing adaptive models, such as early-exit architectures, which allow the inference process to stop early when high confidence is reached, thus dynamically balancing accuracy and efficiency at run-time [4].

- [1] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang. "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge". In: ASPLOS. 2017. DOI: 10.1145/3037697.3037698.
- [2] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang. "Pruning and quantization for deep neural network acceleration: A survey". In: Neurocomputing (2021). DOI: 10.1016/j.neucom.2021.07.045.
- [3] A. Berthelie, T. Chateau, S. Duffner, C. Garcia, and C. Blanc. "Deep Model Compression and Architecture Optimization for Embedded Systems: A Survey". In: JSPS (2020). DOI: 10.1007/s11265-020-01596-1.
- [4] S. Teerapittayanon, B. McDanel, and H.T. Kung. "BranchyNet: Fast inference via early exiting from deep neural networks". In: International Conference on Pattern Recognition (ICPR). 2016. DOI: 10.1109/ICPR.2016.7900006.

PROFIL DES ETUDIANTS SOUHAITE (1 seul choix par projet)

- ☐ AE-SE : spécialité Automatique-Electronique parcours Systèmes Embarqués
- ☒ IR-SI : spécialité Informatique parcours Systèmes Informatiques
- ☐ IR-SC : spécialité Informatique parcours Systèmes Communicants
- ☒ (optionnel) ce projet peut être proposé à un ou des étudiants d'échange sur la partie réalisation seule (semestre 1 et/ou semestre 2)

PRIORITE : 1