Research Grants for PhD students from the China Scholarship Council

Information Form (please read the guidelines carefully on the website www-csc.utt.fr)

Supervisor's name : Kloda | Given names : Tomasz

Status (prof., assistant prof., …) : assistant prof.

Laboratory :	LAAS-CNRS	Website address : https://www.laas.fr/en/
Institution :	Insa de Toulouse	Website address : https://www.insa-toulouse.fr/en/

Scientific competence of the supervisor:

Tomasz Kloda is an Assistant Professor in the Department of Electrical and Computer Engineering at the Insa de Toulouse and LAAS-CNRS. His research focuses on real-time scheduling and embedded systems. Recently he has been working on the integration of AI applications in real-time systems. Before joining Insa de Toulouse, he was at: the Technical University of Munich (Germany), the University of Modena (Italy), and Inria Paris (France). He received two Best Paper Awards.

Two major publications in the field proposed for the PhD :

1. Y. Han, et al.,"Dynamic Neural Networks: A Survey" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 44, no. 11, pp. 7436-7456, 2022. doi: 10.1109/TPAMI.2021.3117837
2. B. Sun, T. Kloda, C. Wu and M. Caccamo, "Partitioned Scheduling and Parallelism Assignment for Real-Time DNN Inference Tasks on Multi-TPU" in 2024 61st ACM/IEEE Design Automation Conference (DAC)

Website address of the personal page : https://homepages.laas.fr/tkloda/

Supervisor's email : tomasz.kloda@insa-toulouse.fr

Description of the research work proposed for a PhD ***Topic # (see list) :*** I-11

Title : Dynamic Neural Networks for Real-Time Systems

Subject :

Deploying neural networks on embedded devices is increasingly challenging due to memory constraints and limited power and processing speed. As certain samples may have very long inference times, meeting their stringent real-time constraints causes hardware oversizing and consumes vast computing power compared to average demand. Dynamic neural networks can opportunistically take advantage of the varying complexity between "easy" and "difficult" inputs, thereby reducing the inference time. Sample-wise dynamic networks adjust model architectures to allocate appropriate computation based on each sample. For instance, in a multi-exit architecture, the execution of intermediate layers can be strategically skipped or conditionally activated. The ability to make early predictions on sample inference time and its criticality level can enable better workload and hardware resource management. Less computation is reserved for samples that are relatively easy to recognize, and immediate deadlines are relaxed for samples that are insignificant from the system's safety perspective. This research will investigate the application of neural network early-exit strategies to real-time systems. Light-weight real-time prediction mechanisms will be proposed to make rapid estimates of the inference time and criticality level based on the current samples. This will allow to construct real-time scheduling algorithms that strike the trade-off between prediction accuracy, execution time, number of input samples and their confidence levels.

Keywords :

AI, real-time systems, neural network optimization

Expected collaborations :

The position will be located in the Southern France at the INSA de Toulouse and LAAS-CNRS, a major research institute in Toulouse, globally renowned for its aeronautics and space industry. The thesis will be supervised by Dr. Tomasz Kloda. The successful candidate will also be encouraged to collaborate with research teams at the University of Rennes 1 (Dr. Angeliki Kritikakou), the INSA de Lyon (Dr. Stefan Duffner) and the Technical University of Munich (Prof. Marco Caccamo).

Background required from the applicant :

Candidates must be current graduate students or have completed their graduate degrees (or equivalent) in Computer Science, Computer Engineering, or Electrical Engineering. Mathematical optimization, machine learning and experience in machine learning frameworks (e.g., TensorFlow, PyTorch, Keras) are a plus.

Existence of a PDF file detailing the proposal ("yes" or "no") : no

(see guidelines on the website www-csc.utt.fr)