Neural Architecture Search framework for efficient and reliable hybrid CNN-Transformer models for Edge AI

Summary

The PhD project aims to develop a neural architecture search (NAS) framework for designing an efficient, robust, and reliable hybrid CNN-Transformer supernetwork capable of generating distinct subnetworks specialised for various HW platforms without extensive retraining. The research objectives include: a) investigating existing and introducing novel training algorithms to enhance the robustness, reliability, and accuracy of the subnetworks within the supernetwork; b) developing fast and efficient search engine algorithms for extracting subnetworks from the trained supernetwork; c) training surrogate predictor models to evaluate key metrics such as accuracy, robustness, reliability, and latency for full-precision and quantised sub-networks. This PhD position is one of the 17 positions in the European Marie Skłodowska-Curie Action Doctoral network "TIRAMISU - Training and Innovation in Reliable and Efficient Chip Design for Edge AI" (2024-2028).

Research field:	Information and communication technology
Supervisors:	Prof. Dr. Maksim Jenihhin Masoud Daneshtalab Prof. Dr. Wolfgang Ecker
Availability:	This position is available.
Offered by:	Tallinn University of Technology School of Information Technologies
Application deadline:	Applications are accepted between September 01, 2024 00:00 and October 31, 2024 23:59 (Europe/Zurich)

Description

The research

Very recently, the landscape of Artificial Intelligence (AI) development has undergone a remarkable transformation, marked by the trend of fusion of AI with the edge computing concept. Artificial intelligence is increasingly being brought to the source of data, i.e., the devices at the network's edge, thus establishing the Edge AI concept. Edge AI is gaining momentum across various industries and services by public authorities because it enables new applications requiring high performance, ultra-low latency with high bandwidth, efficient power use, and intelligence beyond regular computing.

The PhD project aims to develop a neural architecture search (NAS) framework for designing an efficient, robust, and reliable hybrid CNN-Transformer supernetwork capable of generating distinct subnetworks specialized for various Edge AI hardware platforms without extensive retraining. Investigate existing training algorithms and introduce novel training algorithms to enhance the robustness, reliability, and accuracy of the subnetworks within the supernetwork. Developing fast and efficient search engine algorithms for extracting subnetworks from the trained supernetwork. Training surrogate predictor models to evaluate key metrics such as accuracy, robustness, reliability, and latency for full-precision and quantized sub-networks.

This position is a part of new European MSCA Doctoral Network TIRAMISU “Training and Innovation in Reliable and Efficient Chip Design for Edge AI” (2024-2028) https://tiramisu-project.eu/. The action will provide strong interdisciplinary training for future European engineers and researchers driving the innovation for reliable and energy-efficient Edge AI chips. The consortium is strategically designed to foster cross-disciplinary synergies, by seamlessly integrating innovation management research with the technical aspects of Edge AI design. The non-academic sector is represented by a European flagship R&D hub for nanoelectronics - IMEC, a global leader in industrial electronics and the largest semiconductor manufacturer in Germany - Infineon, a trusted automotive solutions provider - Dumarey, the worldwide leader in EDA tools development - Cadence. The academic excellence is established by the top ICT and Technology Innovation engineering universities and Europe's largest application-oriented research organisation - Fraunhofer.

Applicants should fulfil the following requirements:

- (MSCA DN Mobility Rule) **Applicants must not have resided or carried out their main activity (work, studies, etc.) in Estonia for more than 12 months in the 36 months immediately before their date of recruitment.** Compulsory national service, short stays such as holidays, and time spent as part of a procedure for obtaining refugee status under the Geneva Convention are not taken into account. Date of Recruitment means the first day of the employment of the researcher for the purposes of the action (i.e. the starting date indicated in the employment contract or equivalent direct contract).
- a master's degree (or equivalent) degree in Computer Engineering, Computer Science, Artificial intelligence or related areas.
- a clear interest in the topic of the position (candidates with embedded systems and machine learning backgrounds are preferred)
- a good background in linear algebra and math
- a good experience programming in Python
- basic understanding of reliability and machine learning concepts
- English language proficiency
- strong writing and communication skills compatible with an entry-level research position
- capacity to work both as an independent researcher and as part of an international team
- capacity and willingness to provide assistance in organisational tasks relevant to the project

The following experience is beneficial:

- research and/or professional experience, ability and interest to collaborate across disciplines
- familiarity with ML algorithms and DNN architectures
- familiarity with hardware architectures
- familiarity with EDA tools
- previous research publications at conferences or journals

The candidate should submit a research plan for the topic. The candidate can expand on the outlined research scope and propose theoretical lenses to be used.

We offer:

- 4-year PhD position in the Department of Computer Systems that has a sound portfolio of ongoing European and national research projects
- An environment to do excellent research and publications
- Opportunities for training relevant technical and transferable skills aiming academic or industrial careers
- Opportunities for conference visits, research stays and networking with globally leading companies, universities and research centres in the field of research

About the department

The Centre for Trustworthy and Efficient Computing Hardware (TECH) belongs to the Department of Computer Systems. It focuses on cross-layer reliability and self-health awareness technology for tomorrow's complex intelligent autonomous systems and IoT edge devices in Estonia and the EU. The team studies advanced cyber-physical systems characterised by their heterogeneity and emerging computing architectures employing AI-based autonomy. The centre generates knowledge to equip engineers with design-phase solutions and in-field instruments for industry-scale systems to facilitate the system's crashless operation. The core competencies of TECH are: Hardware design; VHDL and Verilog designs; EDA tools (Cadence, Siemens, Synopsys platforms); Application-specific computing platforms (Unmanned Aerial Vehicles); FPGA-based solutions and methodologies; Advanced FPGA SoCs and FPGA development tools (Xilinx Vivado, Altera/Intel Quartus, Lattice Diamond); Software and embedded SW development; Bare-metal and User-space applications; Cross-layer reliability and fault management; ML-based solutions; Functional Safety (ISO26262); Test strategy development and troubleshooting instrumentation; JTAG/IJTAG based solutions; RISC-V processor architectures; DNN hardware accelerators. Head of the centre: Prof. Maksim Jenihhin.
Prof. Dr. Masoud Daneshtalab is Adj. Prof. at TalTech and a full professor at Mälardalen University. His research interests encompass DL acceleration and AutoML. He currently leads seven projects on embedded-friendly AI, focusing on performance, robustness and reliability challenges. He is on the Euromicro board of directors and an associate editor of the Elsevier MICPRO and MDPI images journals. Also, he has supervised over 9 passed PhDs and 3 postdocs, published over 200+ refereed papers, and served on technical program committees of 20+ major AI and design automation conferences.

Prof. Maksim Jenihhin is a tenured associate professor of Computing Systems Reliability at the Department of Computer Systems of Tallinn University of Technology and the head of the research group “Trustworthy and Efficient Computing Hardware”. He received his Ph.D. degree in Computer Engineering from the same university in 2008. His research interests include methodologies and EDA tools for hardware design, verification and debugging as well as nanoelectronics reliability and manufacturing test topics. He supervised 5 PhD theses on these topics and published more than 170 peer-reviewed publications. He is a coordinator for national and European research projects, including Horizon MSCA DN TIRAMISU (2024-2028), Horizon Twinning TAICHIP (2024-2027), H2020 MSCA ITN RESCUE (2017-2021), PRG 2022 CRASHLESS (2022-2027). Prof. Jenihhin is a member of executive and program committees for IEEE ETS, DATE, DDECS, and a number of other international events and served as a guest editor for special issues of journals.
Prof. Dr. Wolfgang Ecker is Distinguished Engineer at Infineon and Professor at Technical University of Munich. His research and innovation focus lies on digital system modeling, digital design automation, SoC architectures, embedded AI and AI for design automation. Wolfgang Ecker published over 200 papers, received six publication awards and has been granted with the German EDA achievement award. He is member of Acatech, the German Academy of Science and Engineering and has been a member of the AI Commission of inquiry of the German Government.
For further information, please contact Masoud Daneshtalab (masoud.daneshtalab@taltech.ee).

To get more information or to apply online, visit https://taltech.glowbase.com/positions/835 or scan the the code on the left with your smartphone.