

Deep Learning Algorithms for Budget Constrained Applications in the IoT Domain

Basic information:

Applications are invited for 15 full-time PhD student positions (called Early Stage Researchers or ESR) for PhD degrees for 36 months. The positions are offered within the Approximate Computing for Power and Energy Optimisation (APROPOS) International Training Network action of the highly appreciated EU-funded Marie Skłodowska Curie grants. The positions are fully funded for a 36-month period. You can choose your position in one of 14 top-level universities and companies in Europe and complement your experience within periods from 3 to 9 months of cross-sector secondments by working with the highly committed industrial partners of the action. You are encouraged to apply for up to three positions within APROPOS network if you believe your profile is suitable for multiple topics. A separate application is needed for each position in case you apply for multiple positions.

APROPOS is an international, multidisciplinary and multi-sectorial training network programme on approximate computing. APROPOS is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No. 956090.

The ESR-specific project is described below. Please read the description carefully before applying. We recommend you choose the position you apply for based on your technical profile and desired technical expertise, rather than based on the hosting location, as all positions involve high-level research mobility.

Project information:

APROPOS will train 15 Early Stage Researchers to tackle the challenges of future embedded and high-performance computing by using disruptive methodologies. Following the current trend, by 2040 computers will need more electricity than the world energy resources can generate. On the communications side, energy consumption in mobile broadband networks is comparable to datacentres. To make things worse, Internet-of-Things will soon connect 20 to 50 billion devices through wireless networks to the cloud.

APROPOS aims at decreasing energy consumption in both distributed computing and communications for cloud-based cyber-physical systems. We propose adaptive Approximate Computing to optimize energy-accuracy trade-offs. Luckily, in many parts of the global data acquisition, transfer, computation, and storage systems there exists the possibility to trade off accuracy to either less power or less time consumed – or both. As examples, numerous sensors are measuring noisy or inexact inputs; the algorithms processing the acquired signals can be stochastic; the applications using the data may be satisfied with an “acceptable” accuracy instead of exact and absolutely correct results; the system may be resilient against occasional errors; and a coarse classification or finding the most probable matches may be enough for a data mining system. By introducing a new dimension, accuracy, to the design optimization, the energy efficiency can even be improved by a factor of 10x-50x.

APROPOS will train the spearheads of the future generation to cope with the technologies, methodologies, and tools for successfully applying Approximate Computing to power and energy saving. The training, in this first ever ITN addressing approximate computing, is to a

large extent done by researching energy-accuracy trade-offs on circuit, architecture, software, and system-level solutions, bringing together world leading experts from European organizations. In addition, we will provide network-wide and local trainings on the substance and on the complementary skills needed in both industrial and academic work life.

Location:

IBM Research - Zurich
AI Automation Group
Säumerstrasse 4,
CH-8803 Rüschlikon, Switzerland

Role description:

AI brings great promise for a new wave of unprecedented productivity across a multitude of industries. Recent AI models have demonstrated extraordinary results in computer vision, natural language processing, as well as many other tasks applied to industrial and engineering domains. Through the years, the complexity of these models has reached enormous levels, with deep learning based models often reaching the hundreds of millions of parameters. In this setting, the traditional approach of human curation and tuning of AI models is not scalable.

The IBM Zurich Research Laboratory is leading the design of innovative automated machine learning and deep learning platforms, targeting enterprises that aim to replace or accelerate traditional human-supervised procedures, with automated services based on AI models. Toward this goal, we are looking to strengthen our team with a highly motivated and skilled scientist that will contribute to the design of such solutions as part of a prestigious European ITN Project – APROPOS.

Given their size and computational complexity, deep learning approaches cannot be easily employed at scale in many real applications. Common issues are indeed execution time, memory footprint, and complexity of long-term model lifecycle maintenance (e.g., re-training models for new incoming data or adapting models to fit new user constraints). Additionally, they generally do not easily fit edge devices, due to their limited power and memory budget. However, with efficient algorithm design we can adapt these complex methods to better scale on limited computing resources. The candidate will develop and evaluate novel algorithms for fast execution of inference models on traditional bare metal machines, on cloud nodes, as well as on edge devices. The research space will cover compression/quantization algorithms, neural network model optimization under user constraints, as well as any other possible close topic that might help us to not only improve memory footprints but also to minimize application run time without sacrificing accuracy. The candidate will also develop efficient edge2cloud execution flows, in a hybrid framework that will enable to process part of the data on edge and the rest on the IBM Cloud. Different data modalities, such as images, text, tabular data, and various sorts of sensor data, provide a challenging setting of applying novel approaches. Henceforth, we adopt our research to deliver tomorrow's AI solutions that follow the needs of our clients. Indeed, the results of the work will be applied to real use cases, in connection with IBM Clients and SME projects in the IoT domain.

The candidate will work full time at the IBM Research – Zurich Laboratory, in the AI Automation group, having the opportunity to work in a unique corporate environment, acquire experience in several areas, publish in top international conferences, learn how to patent innovative ideas, as well as deal with clients on real business cases. The candidate will also spend up to 6 months at Universitat Politecnica de Valencia (UPV), where he/she will also get his PhD degree in Computer Science.

Core activities:

- Develop, standardize, and implement data science and machine learning solutions at scale for edge devices, as well as hybrid cloud systems;
- Create new algorithms for near real-time optimization of models for computer vision, NLP, tabular data, and/or time series, accounting for user/applications constraints on execution time, memory, and power budget, among others;
- Leverage techniques such as quantization, low-precision algorithms, compression, as well as others, to fit the user/application constraints;
- Use the developed models and algorithms inside AI workflows for model lifecycle management, such as (for example) bias mitigation, drift detection, explainability, and constrained optimization;
- Design, develop, and implement proof-of-concepts and prototypes to be ported and included on the IBM Public Hybrid Cloud offering.

Minimum qualifications (mandatory):

- Outstanding university track record, with background in Computing, Machine Learning, Mathematics, Statistics, or equivalent fields;
- 3+ years of proved programming experience in C/C++ and/or Python;
- Proficient in UNIX/Linux;
- Ability to speak and write in English fluently;
- Self-motivated with passion for technology and innovation.

Preferred qualifications:

- Practical experience with Machine Learning and/or Deep Learning frameworks (e.g., scikit-learn, TensorFlow, PyTorch);
- Experience with one or more of the following: web application development, mobile application development, distributed and parallel systems, distributed data storage, information retrieval, computer vision, natural language processing, networking, large software systems development, backend services and REST API development;
- Experience with version control systems and DevOps tools.

Nice to have:

- Contribution to open source projects;
- Proved record of participation in Kaggle competitions (or similar);
- Publications in top AI conferences (NeurIPS, AAAI, etc.);
- Experience with public Cloud environments.

ESR Host Unit: IBM Research - Zurich, Switzerland

Planned secondment(s): The candidate will pursue his/her PhD under the co-supervision of Prof. Enrique S. Quintana-Orti, and upon successful completion of the studies, receive a PhD degree from Universitat Politecnica de Valencia (UPV), Spain. UPV is among the Top 150 universities of the NTU ranking2020 in Computer Science.

Supervisory team:

- Dr. Cristiano Malossi (IBM), main supervisor
- Prof. Enrique S. Quintana-Orti (UPV), academic co-supervisor
- Dr. Florian Scheidegger (IBM), industrial co-supervisor

Enrolment in Doctoral degree: Universitat Politecnica de Valencia (UPV), Spain

Target degree: PhD/D.Sc.(Tech.) in Doctoral Programme Computer Science

Expected start date: April 2021 (M6 from the beginning of the project), in any case not later than September 2021

Approximate gross salary: 5600 CHF/Month

Duration: 36 months

Researcher Profile: First Stage Researcher (R1)

Research Field: Engineering; Computer science; Technology

Type of contract: Temporary

Job Status: Full-time

Hours per week: 40

Trial period: 3 months

IBM Research – Zurich details:

IBM Research – Zurich is one of 12 IBM Research laboratories around the globe. It was established in 1956 and is home to world-class scientists representing more than 45 nationalities. Cutting-edge research and outstanding scientific achievements—most notably two Nobel Prizes—are associated with this lab.

As the largest European branch of IBM Research, its mission – in addition to pursuing innovative research for tomorrow's information technology – is to cultivate close relationships with academic and industrial partners.

IBM Research – Zurich strives to be one of the premier places for top researchers to work, to promote women in IT and science, and to help drive Europe's innovation agenda. A new

facility for collaborative nano-scale research was opened on the IBM Zurich campus in 2011. The Binnig and Rohrer Nanotechnology Center (*top right*) is part of a strategic partnership in nanosciences with ETH Zurich, one of the world's foremost science and engineering universities.

Read more about IBM Research Zurich Laboratory: <https://www.zurich.ibm.com>

Working and living conditions in Switzerland: you can find all information here: <https://www.zh.ch/en/migration-integration/willkommen.html>

Seconding Unit description: Universitat Politècnica de València (UPV, <http://www.upv.es>) is a public university with four campus sites, over 35,000 students, and 2,600 faculty members and research staff. It consists of 44 Departments, most of them in engineering areas. It is the top University in Spain regarding patent production. Contributions to this project will come from the Parallel Architectures group (GAP, <http://ww.gap.upv.es>) of UPV. The group has a 26-year research expertise in different aspects of system architecture, especially on interconnection networks.

Read more about UPV:

<http://www.upv.es/organizacion/la-institucion/documentos/folleto-institucional-upv-eng.pdf>

Eligibility criteria: please check carefully that you are eligible; all conditions below are compulsory

Transnational mobility: The researcher must not have resided or carried out their main activity (work, studies, etc.) in the country of the host organization for more than 12 months in the 3 years immediately prior to the start date. Note: the mobility rule applies to the (main) beneficiary where the researcher is recruited, and not to partners to which the researcher is sent or seconded. It is also only determined at one point in time: at the time when the ESR starts working for APROPOS.

Early-Stage Researchers (ESRs)/fresh MSc graduates: All researchers recruited in APROPOS must have less than 4 years since the completion of their first MSc degree and have not been awarded any doctoral degree at the date of the employment.

Background requirements: The applicant must be in possession of Master of Science (MSc) diploma in a relevant field, such as: electrical engineering, communications engineering, computer engineering, software engineering/computer science, signal processing, radio communications, mathematics, physics, aerospace engineering, mechanical engineering, biomedical engineering, etc.

English language requirements: The candidate must be in possession of an English certificate with good level according to the regulations of the university to which the student will be enrolled for the PhD degree – check the degree requirements of the corresponding doctoral program for details.