Babak Ehteshami Bejnordi

Reseach Scientist • Team Lead • Manager

ehteshami@babakint.com Linkedin babakint.com Google Scholar Amsterdam, The Netherlands

INTERESTS

A research scientist at Qualcomm AI Research (Senior Staff and Manager) where I am a technical team lead focusing on Conditional Computation for efficient deep learning. My primary research focus lies in the realm of efficient Deep Learning for LLMs and Computer Vision. My recent research works have been in the areas of Efficient LLMs, Efficient (Latent) Reasoning, Mixture-of-Experts, Multi-Task Learning, and Continual Learning published as conference papers at NeurIPS, ICML, ICLR, CVPR, ECCV, etc. I am experienced in and enjoy managing people (7y in the industry). I was the organizer of the Qualcomm Innovation Fellowship Program in Europe between 2019-2023. Previously, I obtained my PhD at the Diagnostic Image Analysis Group, Radboud University, the Netherlands, where I worked on the development of ML algorithms for breast cancer diagnostics. During my PhD, I also organized the CAMELYON16 challenge. In 2016, I was a visiting researcher at Harvard University.

PROFESSIONAL EXPERIENCE

QUALCOMM AI RESEARCH
Technical team Lead, Senior Staff Engineer & Manager

Amsterdam, The Netherlands — Nov 2023 – Present

- Team lead focusing on Efficient LLM and VLLM models

QUALCOMM AI RESEARCH
Technical team Lead, Staff Engineer & Manager

Amsterdam, The Netherlands — Nov 2019 – Nov 2023

- Team lead focusing on Conditional Computation for Efficient Deep Learning

QUALCOMM AI RESEARCH
Senior Engineer

Amsterdam, The Netherlands — Apr 2018 – Nov 2019

- Deep Learning Research Scientist

MAPSCAPE B.V.
Deep Learning Engineer

Eindhoven, The Netherlands — Oct 2017 – Mar 2018

- Deep Learning and Computer Vision for Autonomous Driving

EDUCATION

PH.D. IN MACHINE LEARNING & MEDICAL IMAGE ANALYSIS
Radboud University Medical Center, Nijmegen, The Netherlands — Apr 2013 – Jun 2017

VISITING SCHOLAR
Harvard University, Boston, Massachusetts, USA — Jun 2016 – Nov 2016

M.SC. IN ELECTRICAL ENGINEERING
Chalmers University of Technology, Goteborg, Sweden — May 2010 – Dec 2012

B.SC. IN ELECTRICAL ENGINEERING
University of Guilan, Rasht, Guilan, Iran — 2004 – 2008

SELECTED PROJECTS

QUALCOMM AI RESEARCH
Technical Lead focusing on Efficient LLM and VLLM models.

Amsterdam, The Netherlands — May 2023 - Present

- Mixture of Cache-Conditional Experts for Efficient Mobile Device Inference TMLR 2025

- Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design NeurIPS 2024
- Efficient Mixture-of-Experts for mobile devices with limited DRAM NeurIPS 2024 Expo Demo and Under review
- Think Big, Generate Quick: LLM-to-SLM for Fast Autoregressive Decoding ICML 2024 (ES-FoMo II workshop)

QUALCOMM AI RESEARCH — **Technical Lead focusing on Conditional Computation for Efficient DL**

Amsterdam, The Netherlands — Jan 2022 - June 2024

- InterroGate: Learning to Share, Specialize, & Prune Representations for MTL BMVC 2024
- Scalarization for Multi-Task and Multi-Domain Learning at Scale NeurIPS 2023
- MSViT: Dynamic Mixed-Scale Tokenization for Vision Transformers ICCV2023 (NViT workshop)
- Conditional Compute for On-device Video Understanding at NeurIPS Expo Demonstrations
- Revisiting single-gated Mixtures of Experts published at BMVC2022.

QUALCOMM AI RESEARCH — **Efficient Deep Models for Video Processing**

Amsterdam, The Netherlands — Oct 2020 - Dec 2021

- SALISA: Saliency-Based Input Sampling for Efficient Video Object Detection published at ECCV2022.
- FrameExit: Conditional early exiting for efficient video recognition published at CVPR2021 (Oral paper).
- Skip-convolutions for efficient video processing published at CVPR2021.
- Spatio-Temporal Gated Transformers for Efficient Video Processing at NeurIPS2021 (ml4ad workshop).

QUALCOMM AI RESEARCH — **Conditional Computation for Convolutional Neural Networks**

Amsterdam, The Netherlands — May 2018 - Oct 2020

- Conditional Channel Gated Networks for Task-Aware Continual Learning published at CVPR2020 (Oral paper).
- Batch-shaping for learning conditional channel gated networks published at ICLR2020.

HARVARD UNIVERSITY — **Deep learning for Diagnosing Breast Cancer Patients**

Boston, MA, USA — May 2016 - Dec 2016

- Development of a deep learning system for diagnosing breast cancer patients (see publications 2017a and 2017b).
- This work was in Collaboration with NIH, and Mayo Clinic.
- Developed cascade of deep learning models that enables prediction of future invasive breast cancer occurrence among patients which are potentially at high risk of developing breast cancer.

RADBOUD UNIVERSITY — **Lead organizer of CAMELYON16 Machine Learning Challenge**

Nijmegen, The Netherlands — May 2016 - Dec 2016

- The challenge gathered participants from all around the world including Google, Harvard, MIT, etc.
- Performed thorough analysis of Deep Learning algorithms and comparison to expert pathologists.
- Publication at JAMA with 3300+ citations.
- Extensive coverage in over 30 well-known websites and media (e.g. Yahoo News, NOS.nl).
- Highlighted in the White House AI strategic planning report (page 17).

RADBOUD UNIVERSITY — **Machine learning for Breast Cancer Diagnosis**

Nijmegen, The Netherlands — Apr 2013 - Mar 2016

- Co-authored “A survey on deep learning in medical image analysis” published at MEDIA with ∼13,000+ citations.
- Development of context-aware stacked convolutional neural networks to efficiently improve the inclusion of more image context for Whole-slide Image processing published at Journal of Medical Imaging.
- Development of an ML model based on graph theory-based clustering for the detection of pre-invasive cancer (DCIS) in giga-pixel pathology images published at IEEE Transactions in Medical Imaging.
- Development of the first Whole-slide image color standardization published at IEEE Transactions in Medical Imaging.
  - The first algorithm to standardize giga-pixel pathology images (Source code, Executable).
  - Garnered lots of interest among companies and academic institutions and is widely being used by many researchers.
  - Using my algorithm, the team from Harvard & MIT improved its rank from 4th to 1st in CAMELYON16 challenge.

HONORS AND AWARDS

Third highest cited (2000+) work in JAMA 2020

In 2020, with 2000+ citations (currently 3500+), my paper on "Diagnostic assessment of deep learning algorithms lymph node metastases detection" was among the top 3 most cited works of the Journal of American Medical Association (Impact Factor 157) over the past 3 years.

Highest cited survey on Deep Learning for Medical Imaging (∼13,500 citations) 2017

Contributed (third author) to the highest cited survey in deep learning medical imaging to date.

Finalist nomination for Wetenschaps- en Innovatieprijs 2019 2019

Finalist nomination for the Wetenschaps- en Innovatieprijs 2019 of the Federation of Medical Specialists in the Netherlands for my project on artificial intelligence in breast cancer diagnosis.

MedicalPhit Innovation Award of 2016 2016

Won the best MedicalPhit Innovation Award of 2016 in the Netherlands for the study of the use of artificial intelligence in detecting metastases in breast cancer patients.

Awarded research grants 2016

Awarded research grants from BeckLab, Harvard Medical School (2016), Radboud university (2016), Chalmers and Uppsala Universities (2012)

Master's degree with honors 2013

GPA 5.0/5.0 at Chalmers University of Technology.

INTERESTS

Efficient Langauge Modeling, Conditional Computation for Deep Neural Networks, Multi-task Learning, Continual Learning, Efficient Deep Learning and Sparsity, Autonomous Driving, and Medical Imaging.

SKILLS

Programming languages

- Python
- C/C++
- R
- Matlab

Libraries

- Deep Learning: Pytorch, Tensorflow, Keras, Theano
- Computer Vision: OpenCV, scikit-image, scikit-learn, matplotlib

Languages

- English (fluent)
- Persian (native)
- Dutch, Swedish, Italian (basic)

PUBLICATIONS

The full list of my peer-reviewed publications can be found on my google scholar page.

INVITED TALKS

- **GHOST Day:** Efficient Deployment of Large Language Models on Edge Devices at GHOST Day (Applied Machine Learning Conference), Poznan, Poland (2025).
- **Cisco Meraki's GenAI:** Efficient LLM inference on-device (2025).
- **Apple's Efficient LLM reading group:** Efficient inference of Mixture-of-Experts LLMs on-device (2025).
- **DeepLearn 2023 Spring:** Course lectures on Conditional Computation for Efficient Deep Learning at the 9th International School on Deep Learning in Bari, Italy.
- **ELLIS PhD and Postdoc Summit**, Keynote talk at the ELLIS PhD and Postdoc Summit (kick-off program), 2021.
- **TWIML AI:** Podcast interview with Sam Charrington from TWIML AI on Conditional Computation.
- **Broad Institute of MIT and Harvard**, "Practical recommendations for training convolutional neural networks", MIA Seminar: Modeling, Inference, algorithms, 2017, USA, (Video link).

- **Dutch Society for Pattern Recognition and Image Processing**, "Automatic detection of ductal carcinoma in situ in whole slide histopathological images", Eindhoven, The Netherlands, 2015.
- **European Congress on Digital Pathology**, "An algorithm for reducing stain variability in scanned histological slides", Paris, France, 2014.

OTHER ACADEMIC/INDUSTRY EXPERIENCES

- **Qualcomm Innovation Fellowship**: Co-lead in the organization of the Qualcomm Innovation Fellowship (QIF) program in Europe from 2019 to 2023.
- **Deep learning workshop lecturer**: Co-organizer and lecturer at deep learning workshop at Radboud University, the Netherlands (2016).
- **Member of Broad Institute of MIT and Harvard**, Boston, Massachusetts, USA (Jul 2016 – Jul 2017).
- **Supervision and teaching experience:** Supervised more than ∼10 master students of computer science and artificial intelligence for their course/master thesis project. Supervised 5 Ph.D. students for their internship/Qualcomm fellowship projects. I was also a teaching assistant for the course "Computer Aided Diagnosis" at Radboud University between 2013 and 2016.