

Uranik Berisha

PHD RESEARCHER IN EFFICIENT DEEP LEARNING

Stuttgart, Germany | LinkedIn | Google Scholar



I am a PhD researcher at Robert Bosch GmbH and the University of Lübeck, working at the intersection of efficient deep learning, computer vision and model compression. My research explores how neural networks can become smaller, faster and more practical. It has led to publications at CVPR and an Oral presentation at ICCV 2025, as well as two pending patent applications. Alongside research, I enjoy turning technical ideas into useful software. I co-founded Fletza, an AI-powered learning platform, and bring together a multidisciplinary background in mathematics, computer science and software engineering. Completing two bachelor's degrees while working, followed by an accelerated master's degree, reflects the curiosity and persistence I bring to new problems.

Knowledge transfer is a central part of my work. I have presented research to more than 1,000 attendees at ICCV 2025 and regularly prepare and deliver workshops and technical presentations for academic and industry audiences, including at Bosch and the University of Stuttgart.

Timeline



Research

Efficient Data Driven Mixture-of-Expert Extraction from Trained Networks

Authors: Uranik Berisha, Jens Mehnert, Alexandru Paul Condurache
Venue: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)
Abstract: The method converts pretrained Vision Transformers into Mixture-of-Experts models by clustering output activations and extracting the subnetworks responsible for each activation pattern.

Variance-Based Pruning for Accelerating and Compressing Trained Networks

Authors: Uranik Berisha, Jens Mehnert, Alexandru Paul Condurache
Venue: IEEE/CVF International Conference on Computer Vision (ICCV 2025), Oral
Abstract: The method uses activation variance to identify neurons for one-shot structured pruning and reintegrates mean activations to preserve model behavior with minimal fine-tuning.

Projects

Parser Generator

Web-based e-learning environment that makes parser-generator algorithms interactive and easier to understand. Developed as the B.Sc. Computer Science thesis project.
Technologies: JavaScript, HTML, CSS

Plantivity

Gamified productivity app in which each completed focus session plants and sustains a virtual garden, turning progress into visible motivation.
Technologies: Swift, SwiftUI, Core Data, Xcode

ChefBot (2020)

Rule-based AI chatbot that recommends recipes from available ingredients and preferences and passed a small-scale Turing-style evaluation.
Technologies: Python, AIML, Firebase Firestore, Git

open62541 Contributions

Contributed C code to open62541, an open-source OPC UA stack for lightweight clients, servers and embedded applications in industrial automation.
Technologies: C, Git

Talks, Presentations & Workshops

ICCV 2025 Oral Presentation

Audience of 1,000+

Presented Variance-Based Pruning for Accelerating and Compressing Trained Networks to more than 1,000 attendees, communicating the motivation, methodology and practical relevance of the research.

Generative AI, Machine Learning & Deep Learning Workshop

University of Stuttgart

Introduced core concepts with an emphasis on practical relevance. Covered next-token prediction, GPT architecture and training, data collection and filtering, pretraining, instruction tuning, dialogue optimization and learning from user feedback.

Expertise & Skills

Research expertise in machine learning and deep learning, computer vision and Vision Transformers, Mixture-of-Experts architectures, and model pruning and compression.

Programming Languages

Python	■■■■■
JavaScript / TypeScript	■■■■□
SQL	■■■■■
C/C++	■■■■□
Java	■■■□□
Swift	■■■■□

Skills & Technologies

PyTorch	■■■■■
NumPy / Pandas / SciPy / scikit-learn	■■■■□
React / Node.js	■■■■□
Git / Linux	■■■■■
CUDA	■■□□□

Languages

German	■■■■■
English	■■■■■
Swedish	■■■■■
Albanian	■■■■■
Spanish	■■■□□