Karlsruhe Institute of Technology

Job Offer

Efficient Long-Context Large Models for Vision and VLA Systems

For more than a decade the Institute of Measurement and Control Systems (MRT) has been one of Germany's leading autonomous driving labs. Its unique advantage is the experimental vehicles with sensors, actuators and a complete software stack for automated urban driving. The existing equipment and experience not only enable continuous cutting-edge research in all areas of autonomous driving, but also successful participation in scientific competitions and a wide range of collaborations with leading industrial partners. In addition, MRT publishes datasets such as the KITTI Vision Benchmark Suite or the KITScenes datasets, which influence autonomous driving research.

Tasks

Large foundation models and vision-language-action models increasingly rely on long context windows to reason over high-resolution images, videos, multi-camera sensor streams, maps, interaction histories and language instructions. However, standard attention mechanisms scale unfavorably with context length, and large models remain difficult to deploy under real-time, memory and energy constraints in autonomous systems. This role will investigate efficient long-context large models for computer vision and VLA systems in the application context of automated driving. The research will combine attention optimization, token reduction, model pruning and quantization to make large models scalable while preserving accuracy, robustness and downstream action quality. The thesis will build on established ideas from Transformers [1], long-sequence attention [2], memory-efficient attention [3], vision transformers [4], deep compression [5] and recent VLA systems [6]. You may focus on:

+ Extending large vision and vision-language-action models to long context windows for video, multi-camera perception and embodied decision making.
+ Optimizing attention mechanisms for high-resolution visual tokens, temporal memory and multi-modal context with efficient GPU implementation and real-time inference.
+ Developing pruning, quantization and token selection techniques that reduce latency, memory and energy consumption while preserving model performance.
+ Evaluating the developed methods on computer vision benchmarks, VLA tasks and real-world autonomous driving or robotic scenarios.

We offer

An exciting working and research environment with

+ an interdisciplinary working environment with self-driving experimental vehicles (BMW 7 and Mercedes Maybach) and leading partners from science and industry
+ a scientific challenge that enables publications at top-tier conferences for machine learning, computer vision, robotics and automated driving (NeurIPS, ICLR, ICML, CVPR, ICCV, ECCV, ICRA, IROS, IV)
+ collaboration in a strong and well-recognized team
+ potential participation in the excellent Karlsruhe School of Optics and Photonics (KSOP)
+ opportunity to earn a PhD at KIT and
+ compensation according to a full TV-L E13 position

We are looking for

A dedicated and creative person with

+ an excellent Master's degree in engineering, computer science, robotics or similar
+ knowledge of deep learning and large model architectures, ideally transformers, attention mechanisms, computer vision or robotics
+ experience in software development (Python/C++; PyTorch, JAX or TensorFlow) under Linux/ROS
+ interest in efficient inference, GPU programming, model pruning, quantization or deployment on embedded hardware
+ interest in working independently and scientifically, and
+ eagerness to actively participate in a great team.

Application

Please apply with **Prof. Dr.-Ing. Christoph Stiller** (stiller@kit.edu) and include CV, transcript of records as well as — optionally — code samples (GitHub) or publications. We aim to balance the number of employees (f/m/d). Therefore, we particularly ask female applicants to apply for this job. Recognized severely disabled persons will be preferred if they are equally qualified.

[1]: Vaswani et al., Attention Is All You Need, NeurIPS 2017. [2]: Beltagy et al., Longformer: The Long-Document Transformer, 2020. [3]: Dao et al., FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, NeurIPS 2022. [4]: Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021. [5]: Han et al., Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, ICLR 2016. [6]: Brohan et al., RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, CoRL 2023.