Professorship Big Geospatial Management
TUM School of Engineering and Design
Technichal University of München

Master Thesis

Efficient Retrieval Augmented Generation on Edge Devices

Retrieval Augmented Generation (RAG) combines information retrieval with large language models to enhance text generation with domain-specific knowledge. Traditional RAG systems operate in cloud environments with abundant resources but growing demands for privacy and reduced latency require deploying RAG on edge devices. Edge deployment faces severe memory, computational, and power constraints, as current RAG implementations require several gigabytes of memory. Recent advances in model compression techniques like QLoRA (Quantized Low-Rank Adaptation) and specialized hardware like NVIDIA Jetson platforms offer promising solutions.

Location: Ottobrunn/Munich/Remote

Duration: 3 to 6 months depending on your study program

Your topic:

The primary challenge is developing an RAG system that maintains acceptable performance within edge device constraints. This requires balancing retrieval accuracy, generation quality, real-time inference speeds, and memory footprint. Key technical challenges include optimizing embedding models for efficient similarity search, compressing language models while preserving reasoning capabilities, developing memory-efficient retrieval indices, and coordinating the pipeline for minimal latency. You are required to design, implement and evaluate an RAG system optimized for NVIDIA Jetson edge devices. Concretely, the thesis is expected to include

- Developing an efficient RAG pipeline (potentially with retrieval index and memory management optimization).
- Deploying the system on NVIDIA Jetson with TensorRT optimization to achieve acceptable inference throughput.
- Establishing benchmarks measuring retrieval accuracy, generation quality, latency, memory usage, and power consumption.
- (optionally) Apply quantization (INT8/INT4) and QLoRA fine-tuning to embedding and language models, evaluating trade-offs between size, accuracy, and inference speed.

Related Work:

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
- Izacard, G., & Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.
- Tambe, T., Hooper, C., Pentecost, L., Jia, T., Yang, E. Y., Donato, M., ... & Wei, G. Y. (2021, October). Edgebert: Sentence-level energy optimizations for latency-aware multi-task nlp inference. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 830-844).
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115.

Qualifications:

- Programming & libraries: Python, PyTorch/TensorFlow, (preferably TensorRT, CUDA)
- Fundamental knowledge and experience with language modeling
- Interest and experience in literature-based work with good scientific practice
- Enrolled full time student within Computer Science, Electrical Engineering, GIS or similar field of study
- Fluent English is mandatory; German would be an asset

Applications via Mail with CV and transcript to:

Advisor: M.Sc. Xuanshu Luo
Raum: 9377.01.113
Telefon: 089/289-555 51
Email: xuanshu.luo@tum.de