Forschungsprojektarbeit im Profil Data Science

The Real Cost of Privacy: Quantifying the Trade-offs for On-Device Vision-Language Models

Initial Position

Modern AI, especially Vision-Language Models (VLMs), relies heavily on centralized cloud infrastructure. This approach compromises user privacy since data must be sent to a server and increases energy consumption. This thesis investigates the essential trade-off: **how much accuracy is sacrificed to gain total privacy and significant energy savings?** Your role is to quantify this cost-benefit analysis by building, deploying, and benchmarking these new efficient models.

Objectives and Procedure

The overarching objective is to quantify the real-world trade-offs of using on-device VLMs versus cloud models across four metrics: Accuracy, Latency, Model Size, and Energy Consumption. This is achieved through three progressive, practical projects:

Project 1: Establish the data foundation for the thesis by curating a novel, privacy-focused Visual Question Answering (VQA) benchmark. This includes defining the "gold standard" for answer quality using a high-fidelity cloud VLM. VQA is used due to its high real-world impact in enabling intuitive visual AI, and its challenging multimodal nature.

Project 2: Develop the core automated evaluation tool—the "LLM-as-a-Judge"—to reliably score answer quality. This enables an initial analysis of optimization methods by quantifying the accuracy loss from model quantization on a server.

Project 3 (Master thesis): Conduct the definitive experimental analysis by deploying efficient VLMs onto a physical edge device and measure the trade-off between the model's accuracy, speed, and actual energy consumption.

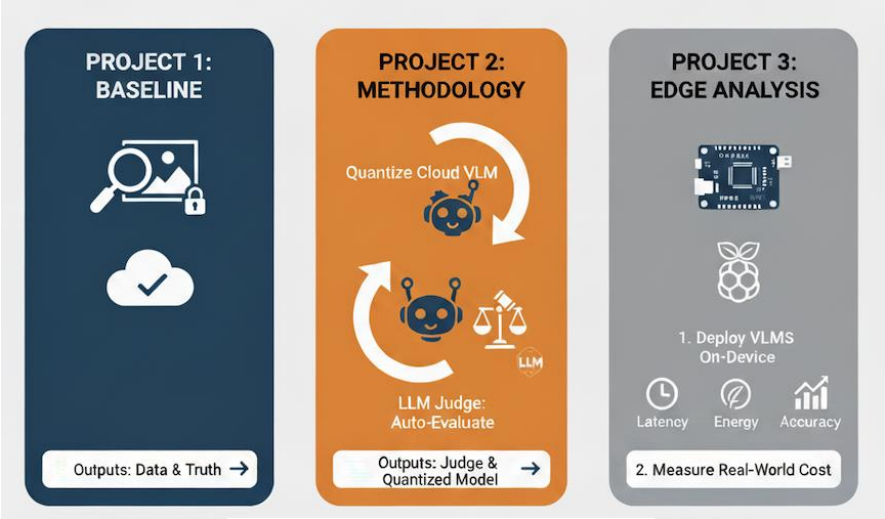

Figure 1: Envisioned Thesis Flow

Required skills: Programming skills, Basic DS skills, General familiarity with LLMs, curious and proactive learner

References

Visual Instruction Tuning, Liu et al. (2023)
LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model, Zhu et al. (2024)
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, Zheng et al (2023)

Studienart: [x] Vollzeitstudium
[x] Teilzeitstudium 50% mit Assistenzanstellung

Projektorganisation: Arbeit in einem Projektteam / Einzelarbeit. **Projektfinanzierung:** Free research

Arbeitsort: Windisch **Advisor**: Prof. Dr. Jelena Milosevic