Bachelorarbeit

Thema: Large Language Models zur Onboard-Verarbeitung von natürlicher Sprache in der mobilen Robotik

Beschreibung:

Die Integration von Large Language Models (LLMs) in die mobile Robotik eröffnet viele Möglichkeiten für diverse Anwendungen. Durch die Anbindung eines LLMs kann ein Roboter nicht nur für Konversationen genutzt werden, sondern auch dazu dienen, den Roboter mittels natürlicher Sprache zu steuern. Insbesondere LLMs wie GPT-4 verfügen über ein beeindruckendes Weltwissen, das hierfür als implizite Wissensgrundlage genutzt werden kann.
Aus verschiedenen Gründen, darunter das Fehlen einer stabilen Internetverbindung, Datenschutzproblematiken und die Abhängigkeit von externen Diensten, besteht jedoch das Interesse, LLMs lokal auf der Onboard-Hardware eines mobilen Roboters auszuführen. Hierfür existieren bereits Verfahren [1-5], die grundsätzlich ohne ressourcenintensives Training genutzt werden können. Darüber hinaus stehen optimierte Inferenzbeschleuniger [6-8] zur Verfügung, die das Potenzial besitzen, LLMs direkt auf einem Roboter auszuführen.

Das Ziel dieser Bachelorarbeit besteht darin, verschiedene geeignete LLMs hinsichtlich ihrer Eignung für Konversationen sowie der Steuerung des Roboters mittels natürlicher Sprache zu untersuchen. Hierbei soll ein geeignetes vortrainiertes Verfahren identifiziert und unter Verwendung verschiedener Inferenzbeschleuniger auf hardwaretypischer Roboterplattform [9] umfassend getestet werden, und an die am FG NI&KR verwendete Middleware MIRA integriert werden.

Detaillierte Aufgabenstellung:

- systematische Aufarbeitung des State of the Art zum betrachteten Themenfeld ausgehend von [1-5]
- Auswahl eines oder mehrerer geeigneten Large Language Models zum Ermöglichen von Konversation und dem Auslösen von Steuerkommandos durch JSON konforme Ausgaben
- Detaillierte Betrachtung von verschiedenen Inferenzlösungen für LLMs ausgehend von [6-8] mit detaillierter Betrachtung von Laufzeiteigenschaften für typische Hardware eines mobilen Roboters des FG NI&KR [9] (Intel NUC11PHKi7 und NVIDIA Jetson AGX Orin 64GB)
- Anbindung eines Verfahrens an die am Fachgebiet verwendete Middleware MIRA
- Ausarbeitung von Präsentationen für den Eröffnungs-, und Abschlussvortrag
- Anfertigen der Abschlussarbeit entsprechend der Vorgaben des FG NI&KR

Notwendige Voraussetzungen:

- Abschluss der Vorlesungen „Neuroinformatik“ und „Deep Learning for Computer Vision“
- gute Kenntnisse im Bereich Deep Learning mit Vorwissen zu LLMs
- gutes mathematisches Verständnis und Erfahrung in der Programmierung mit Python

Literatur:

[1] Touvron, et al.: LLaMA: Open and Efficient Foundation Language Models, arXiv, 2023.
[2] Touvron, et al.: LLaMA 2: Open Foundation and Fine-Tuned Chat Models, arXiv, 2023.
[3] Jiang, et al.: Mistral 7B, arXiv, 2023
[4] Wang, et al.: Self-Instruct: Aligning Language Model with Self Generated Instructions, ACL, 2023
[5] Hu, et al.: LoRA: Low-Rank Adaptation of Large Language Models, ICLR, 2022
[6] NVIDIA: TensorRT-LLM, GitHub, 2023
[7] Kwon et al.: Efficient Memory Management for Large Language Model Serving with PagedAttention, SOSP, 2023
[8] Gerganov , et al.: llama.cpp, GitHub, 2023
[9] Fischedick, et al.: Bridging Distance with a Collaborative Telepresence Robot for Older Adults – Report on Progress in the CO-HUMANICS Project, ISR, 2023

- Google Scholar scholar.google.com
- Proceedings der rel. Konferenzen (IROS, ICRA, NIPS, ICML, ICLR, IJCNN, WCCI, ICANN, CVPR, ICCV, ECCV, BMVC, ICPR, ICIP, …)

Betreuer: Söhnke B. Fischedick, M.Sc. (soehnke.fischedick@tu-ilmenau.de)
Betr. Hochschullehrer: Prof. Dr. H.M. Groß
Bearbeiter: **Simon Pfannschmidt**