Technische Hochschule Nürnberg

BA/MA-Abschlussarbeit

TinyAI Experiment: Implementierung von neuronalen Netzen und LLM auf einem tragbaren Linux-Minicomputer

Kurzbeschreibung

Aufgrund ihrer inhärenten Parallelität haben sich neuronale Netze als ein sehr starkes KI-System erwiesen, das nicht nur die neuesten Entwicklungen bei großen Sprachmodellen antreibt, sondern auch ein breites Spektrum an industriellen Anwendungen. In den letzten Jahren hat sich der Schwerpunkt jedoch von großen neuronalen Modellen und stromhungrigen Anwendungen in Rechenzentren auf eingebettete und sogar sehr kleine neuronale Netzsysteme verschoben. Dieser Trend wird durch die Verbreitung eingebetteter Unterhaltungselektroniksysteme wie Fitness-Tracker, tragbare Spielgeräte usw. unterstützt. Um den Ressourcen- und Energiebedarf zu minimieren, konzentrieren sich die jüngsten Entwicklungen in der TinyAI konzentrieren sich darauf, große neuronale Netzwerke zu komprimieren und zu quantisieren, damit sie in kleine, ressourcenbeschränkte Geräte passen. Die theoretische Arbeit, die sowohl von der Industrie als auch von der akademischen Welt geleistet wird, befasst sich mit Methoden, um große Modelle mit einem angemessenen Verlust an Leistung und Genauigkeit zu verkleinern. Hier spielen einfache Implementierungen und Effizienz eine große Rolle.
In diesem Projekt geht es darum, den Einsatz neuronaler Netze auf einem kleinen Handheld-Linux-Minicomputer zu untersuchen. Der Ben NanoNote ist ein Taschencomputer mit dem Linux-basierten OpenWrt-Betriebssystem. Das von Qi Hardware entwickelte Open-Source-Software- und -Hardware-Gerät wurde bereits als „der kleinste Linux-Laptop der Welt“ bezeichnet. Das System verfügt über eine 336 MHz XBurst JZ4720 MIPS CPU, 32 MB SDRAM und 2 GB Flash-Speicher. Das Ziel des Projekts ist es, ein einfaches neuronales Netzwerk für Regression und Klassifizierung zu entwickeln und einzusetzen. Die entwickelte Pipeline für den Einsatz des neuronalen Netzes wird dann verwendet, um die Barebone-Referenzimplementierung von GPTv2 zu portieren, die von llm.c in der Sprache C angeboten wird.

Tasks

- Einführung in die Anatomie der neuronalen Netze für Regression und Klassifikation.
- Einführung in llm.c und die Referenzimplementierung des GPTv2-Modells.
- Vertrautmachen mit der Handheld Linux Minicomputer Umgebung.
- Implementierung der einfachen neuronalen Netze für Regression und Klassifikation.
- Implementierung von llm.c und Cross-Kompilierung für den Handheld Linux Minicomputer.
- Einsatz auf dem Handheld-Linux-Minicomputer und Leistungsanalyse.

Voraussetzungen

- Erfahrung mit Neuronale Netzwerke und/oder LLM
- Gute Embedded Systems Kenntnisse
- Gute Linux Kenntnisse
- Sehr gute Programmierkenntnisse (C/C++, Python)

Betreuer

Prof. Dr. Ing. Cristian Axenie, M.Sc.