Technische Hochschule Nürnberg

BA/MA-Abschlussarbeit

TinyAI Experiment: Implementierung von neuronalen Netzen und 1-bit LLM auf einem 9$ Micro-Computer

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 Linux-Minicomputer zu untersuchen. C.H.I.P. ist ein günstiger Single Board Computer, der von Next Thing Co. entwickelt wurde. Er kostet nur 9 US-Dollar und verfügt über einen 1 GHz ARMv7-Prozessor, 512 MB Arbeitsspeicher und 3,8 GB (etwa 4 GB) Speicherplatz, auf dem ein angepasstes, leichtes Debian-Betriebssystem läuft. Das Gerät verfügt außerdem über integriertes Wi-Fi und Bluetooth, was es zu einer vielseitigen Option für den Einsatz als Standalone-Computer macht, zumal der Chip auch einen ARM Mali-400-Grafikprozessor und einen H263-, H264- und vp8-Hardware-Video-Decoder enthält. 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 1-bit LLM basiert auf T-MAC zu implementieren. T-MAC ist eine Kernel-Bibliothek, die die Multiplikation von Matrizen mit gemischter Genauigkeit (int1/2/3/4 x int8/fp16/fp32) direkt unterstützt, ohne dass eine Dequantisierung mit Hilfe von Lookup-Tabellen erforderlich ist.

Tasks

- Einführung in die Anatomie der neuronalen Netze für Regression und Klassifikation.
- Einführung in 1-bit LLM und die Referenzimplementierung des T-MAC-Modells.
- Vertrautmachen mit der CHIP Linux Minicomputer Umgebung.
- Implementierung der einfachen neuronalen Netze für Regression und Klassifikation.
- Implementierung von 1-bit LLM T-MAC und Cross-Kompilierung für den Linux Minicomputer.
- Einsatz auf dem CHIP-Linux-Minicomputer und Leistungsanalyse.

Voraussetzungen

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

Betreuer

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