Energiespar-Beschleunigungschips für neuronale Netzwerke: Fortschritte bei benutzerdefinierten Gleitkommatechniken

Kurzfassung

Diese Dissertation ist eine Untersuchung zu Designtechniken, die eine benutzerdefinierte Gleitkommaberechnung (FP) für stromsparende neuronale Netzwerkbeschleuniger in ressourcenbeschränkten eingebetteten Systemen beinhalten. Sie konzentriert sich auf die Nachhaltigkeit der zukünftigen Omnipräsenz künstlicher Intelligenz (KI) durch die Entwicklung effizienter Hardware-Engines und betont das Gleichgewicht zwischen energieeffizienten Berechnungen, Inferenzqualität, Anwendungsvielfalt und plattformübergreifender Kompatibilität.

Die Studie betont die Rolle von Machine Learning (ML) bei der Weiterentwicklung des stromsparenden Internets der Dinge (IoT), insbesondere in der Industrie 4.0, unter Berücksichtigung der rechnerischen und energetischen Herausforderungen, die von ML-Algorithmen ausgehen. Die Dissertation zielt darauf ab, die Hardwareeffizienz zu verbessern, vor allem durch approximative Rechentechniken wie Quantisierung. Dieser Ansatz verbessert zwar die Energieeffizienz und beschleunigt die Berechnungen, erfordert jedoch ein sorgfältiges Design, um die numerische Präzision mit der Modellgenauigkeit in Einklang zu bringen.

Die Forschung stellt eine Hardware-Design-Methodik für Low-Power-Inferenz von neuronalen Spike-by-Spike (SbS) Netzwerken vor. Trotz der reduzierten Komplexität und Rauschrobustheit von SbS-Netzwerken bleibt ihr Einsatz in eingeschränkten eingebetteten Geräten aufgrund der hohen Speicher- und Rechenkosten eine Herausforderung. Die Dissertation schlägt ein neuartiges Multiply-Accumulate (MAC) -Hardwaremodul vor, das das Gleichgewicht zwischen Rechengenauigkeit und Ressourceneffizienz in FP-Operationen optimiert. Dieses Modul verwendet einen hybriden Ansatz, der Standard-FP mit benutzerdefinierten 8-Bit-FP- und 4-Bit-logarithmischen numerischen Darstellungen kombiniert, wodurch eine Anpassung basierend auf anwendungsspezifischen Einschränkungen ermöglicht wird und erstmals eine Beschleunigung in eingebetteten Systemen implementiert wird.

Darüber hinaus stellt die Studie ein Hardware-Design für Low-Power-Inferenz in Convolutional Neural Networks (CNNs) vor, das auf Sensor-Analyse-Anwendungen abzielt. Dies schlägt ein Quantisierungsschema für Hybrid-Float6 (HF6) und einen dedizierten Hardwarebeschleuniger vor. Die vorgeschlagene Quantization-Aware Training (QAT) -Methode zeigt trotz der numerischen Quantisierung eine verbesserte Qualität. Das Design stellt die Kompatibilität zu Standard-ML-Frameworks wie TensorFlow Lite sicher und unterstreicht sein Potenzial für den praktischen Einsatz in realen Anwendungen.

Zusammenfassend befasst sich diese Dissertation mit der kritischen Herausforderung, Rechengenauigkeit mit Energieeffizienz in KI-Hardware-Engines mit Inferenzqualität, Anwendungsvielfalt und plattformübergreifender Kompatibilität als Designphilosophie zu harmonisieren. Sie trägt wesentlich zur Entwicklung nachhaltiger neuronaler Netzwerkprozessoren bei, die für die zunehmend vernetzte und intelligente Welt von entscheidender Bedeutung sind.