Low-Power Neural Network Accelerators: Advancements in Custom Floating-Point Techniques

Abstract

This dissertation presents an investigation into design techniques involving custom floating-point computation for low-power neural network accelerators in resource-constrained embedded systems. It focuses on the sustainability of AI transformation through the development of efficient hardware engines, emphasizing the balance between energy-efficient computations, precision, reliability, and cross-platform integration.

The study emphasizes the role of Machine Learning (ML) in advancing IoT, particularly in Industry 4.0, and acknowledges the computational and energy challenges posed by ML algorithms. The dissertation aims to enhance hardware efficiency, primarily through approximate computing techniques like quantization. This approach, while improving energy efficiency and accelerating computations, requires careful design to balance precision and model accuracy.

Specifically, the research presents a hardware design methodology for low-power inference of Spike-by-Spike (SbS) neural networks. Despite the reduced complexity and noise robustness of SbS networks, their deployment in constrained devices is challenging due to high memory and computational costs. The dissertation proposes a novel Multiply-Accumulate (MAC) module that optimizes the balance between computational accuracy and resource efficiency in Floating-Point (FP) operations. This module employs a hybrid approach, combining standard FP with custom 8-bit FP and 4-bit logarithmic number representations, allowing for customization based on application-specific constraints, enabling deployment on embedded systems.

Additionally, the study introduces a hardware design for low-power inference in Convolutional Neural Networks (CNNs), targeting sensor analytics applications. This proposes a Hybrid-Float6 (HF6) quantization scheme and a dedicated hardware accelerator. The proposed Quantization-Aware Training (QAT) method demonstrates improved precision despite quantization. The design ensures compatibility with standard ML frameworks as TensorFlow Lite, highlighting its potential for practical deployment in real worlds embedded applications.

In summary, this dissertation addresses the critical challenge of harmonizing computational accuracy with energy efficiency in AI hardware design. It contributes significantly to the development of sustainable neural network processors, crucial for the increasingly connected and intelligent world.

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.