Low-Power Neural Network Accelerators: Advancements in Custom Floating-Point Techniques (One Page Short Version)

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 präsentiert eine Untersuchung von Entwurfstechniken, die benutzerdefinierte Gleitkomma-Berechnungen für energieeffiziente neuronale Netzwerkbeschleuniger in ressourcenbeschränkten eingebetteten Systemen beinhalten. Sie konzentriert sich auf die Nachhaltigkeit der KI-Transformation durch die Entwicklung effizienter Hardware-Engines und betont das Gleichgewicht zwischen energieeffizienten Berechnungen, Präzision, Zuverlässigkeit und plattformübergreifender Integration.

Die Studie hebt die Rolle des maschinellen Lernens (ML) beim Fortschritt des Internets der Dinge (IoT), insbesondere in der Industrie 4.0, hervor und erkennt die rechnerischen und energetischen Herausforderungen, die durch ML-Algorithmen entstehen. Das Ziel der Dissertation ist es, die Effizienz der Hardware zu steigern, vor allem durch Näherungsrechentechniken wie Quantisierung. Dieser Ansatz verbessert zwar die Energieeffizienz und beschleunigt die Berechnungen, erfordert jedoch eine sorgfältige Gestaltung, um Präzision und Modellgenauigkeit auszugleichen.

Insbesondere präsentiert die Forschung eine Hardware-Entwurfsmethodik für energieeffiziente Inferenz von Spike-by-Spike (SbS) neuronalen Netzwerken. Trotz der reduzierten Komplexität und Geräuschrobustheit von SbS-Netzwerken ist deren Einsatz in ressourcenbeschränkten Geräten aufgrund hoher Speicher- und Rechenanforderungen herausfordernd. Die Dissertation schlägt ein neuartiges Multiply-Accumulate (MAC) Modul vor, das das Gleichgewicht zwischen rechnerischer Genauigkeit und Ressourceneffizienz in Gleitkommaoperationen (FP) optimiert. Dieses Modul verwendet einen hybriden Ansatz, der standardmäßige FP mit benutzerdefinierten 8-Bit-FP und 4-Bit-logarithmischen Zahlendarstellungen kombiniert, was eine Anpassung basierend auf den spezifischen Anforderungen der Anwendung ermöglicht und den Einsatz in eingebetteten Systemen erleichtert.

Darüber hinaus führt die Studie ein Hardware-Design für energieeffiziente Inferenz in Convolutional Neural Networks (CNNs) ein, das auf Sensoranalytikanwendungen abzielt. Dies schlägt ein Hybrid-Float6 (HF6) Quantisierungsschema und einen speziellen Hardware-Beschleuniger vor. Die vorgeschlagene quantisierungsbewusste Schulungsmethode (QAT) demonstriert verbesserte Präzision trotz Quantisierung. Das Design gewährleistet die Kompatibilität mit standardisierten ML-Frameworks wie TensorFlow Lite und hebt das Potenzial für den praktischen Einsatz in realen eingebetteten Anwendungen hervor.

Zusammenfassend adressiert diese Dissertation die kritische Herausforderung, rechnerische Genauigkeit mit Energieeffizienz im KI-Hardware-Design in Einklang zu bringen. Sie leistet einen signifikanten Beitrag zur Entwicklung nachhaltiger neuronaler Netzwerkprozessoren, die für die zunehmend vernetzte und intelligente Welt von morgen entscheidend sind.