Yarib Israel Nevarez Esparza
Low-Power Neural Network Accelerators: Advancements in Custom Floating-Point Techniques

December 9, 2023

Abstract

The expansion of AI is addressing a new era characterized by omnipresent connected devices. To ensure the sustainability of this transformation, it is imperative to adopt design strategies that harmonize precise computational results with economically viable system architectures. Consequently, refining the efficiency and quality of AI hardware engines stands as a critical consideration in this evolution. This necessitates a balanced approach that prioritizes energy-efficient computations, precise and reliable results, and integration across various platforms and devices.

ML algorithms are serving as the foundational enabler for the integration of AI into IoT devices, particularly in the context of Industry 4.0. These advancements are shaping applications to be more intelligent and economically rewarding. This transformation improves numerous domains, from scientific research to industrial processes and everyday living. However, this technological evolution also brings its own set of challenges. ML algorithms pose significant computational and energy demands. Consequently, a central objective of this dissertation is to explore innovative methods for enhancing the hardware efficiency of computing engines.

Approximate computing techniques, such as quantization, exploit the inherent error resilience of ML algorithms to address key design concerns in computer systems: energy efficiency, performance, and chip area. Quantization, which involves reducing the number of bits used to represent numbers, can significantly lower power consumption and data movement, thereby enhancing energy efficiency by employing compact arithmetic units that save chip area. These techniques often yield computation acceleration due to reduced data sizes, which promotes faster, more parallel, and pipelined processing, particularly in neural network computation. However, this approach introduces a trade-off between precision and model accuracy, necessitating proper hardware design methodologies. While state-of-the-art methods are advancing, significant research opportunities remain, especially for accelerators with custom FP computation.

In this dissertation, a hardware design methodology is presented for low-power inference of SBS neural networks for embedded applications, within the field of SNNs. Compared to conventional SNNs employing the LIF mechanism, SBS neural networks are highlighted for their reduced model complexity and exceptional noise robustness. However, despite their advantages, SBS networks inherently possess a memory footprint and computational cost that makes them challenging for deployment in constrained devices. To solve this issue, this research leverages the intrinsic error resilience of SBS models, aiming to enhance performance and reduce hardware complexity, while avoiding quantization. Specifically, this research introduces a novel MAC module designed to optimize the balance between computational accuracy and resource efficiency of FP operations. This MAC module features configurable quality through a hybrid approach. It combines standard FP number representations with a custom 8-bit FP format, as well as a 4-bit logarithmic number representation. This design excludes the use of a sign bit, further contributing to the compact and efficient representation of numbers. This design enables the MAC module to be tailored to the specific resource constraints and performance requirements of a given application, making SBS neural networks possible for deployment in resource-constrained environments.

In the field of CNNs, this dissertation presents a hardware design methodology for low-power inference, specifically targeting sensor analytics applications. Central to this work is the proposal of the HF6 quantization scheme and its dedicated hardware accelerator, designed to function as a Conv2D TP. This quantization strategy employs a hybrid number representation, combining standard FP and a 6-bit FP format. This strategy allows for a highly optimized FP MAC, reducing mantissa multiplication into a multiplexer-adder operation. This research introduces a QAT method that, in certain cases, offers beneficial regularization effects. The efficacy of this exploration is demonstrated with a regression model, which improves its precision despite the applied quantization. For ML portability, the custom FP representation is encapsulated within a standard format – a design characteristic that enables the proposed hardware to process it automatically. To validate the interoperability of this approach, the hardware architecture is integrated with TensorFlow Lite, demonstrating compatibility with industry-standard ML frameworks and affirming the potential for practical deployment in various sensing applications while maintaining compliance with established ML infrastructure.

This dissertation addresses an essential challenge in the current technological landscape: the harmonization of computational accuracy with energy efficiency and compatibility of hardware solutions. This dissertation stands as a significant contribution towards the development of a sustainable next-generation of neural network processors, essential to empower the increasingly connected and intelligent world of tomorrow.

Kurzfassung

Die Ausweitung Künstlicher Intelligenz (KI) führt in eine neue Ära, die von omnipräsent vernetzten Geräten geprägt ist. Um die Nachhaltigkeit dieses Wandels zu gewährleisten, ist es unerlässlich, Designstrategien zu verfolgen, die präzise Rechenergebnisse mit wirtschaftlich tragfähigen Systemarchitekturen in Einklang bringen. Daher ist die Verfeinerung der Effizienz und Qualität von KI-Hardware-Engines bei dieser Entwicklung von entscheidender Bedeutung. Dies erfordert einen ausgewogenen Ansatz, der die Energieeffizienz der Berechnungen, Präzision und Zuverlässigkeit der Ergebnisse sowie die Integration über verschiedene Plattformen und Geräte hinweg priorisiert.

Machine Learning (ML)-Algorithmen dienen als grundlegende Voraussetzung für die Integration von KI in Geräte des Internets der Dinge (IoT), insbesondere im Kontext von Industrie 4.0. Diese Weiterentwicklungen beeinflussen die Gestaltung von Anwendungen, die intelligenter und ökonomisch vorteilhafter werden sollen. Dieser Wandel verbessert zahlreiche Bereiche, von der wissenschaftlichen Forschung über industrielle Prozesse bis hin zum Alltag. Allerdings bringt diese technologische Entwicklung auch eigene Herausforderungen mit sich. ML-Algorithmen sind mit einem erheblichen Rechen- und Energiebedarf verbunden. Zentrales Ziel dieser Dissertation ist es daher, innovative Methoden zur Verbesserung der Hardwareeffizienz von Rechenmaschinen zu erforschen.

Approximative Rechentechniken, wie die Quantisierung, nutzen die inhärente Fehlerresistenz von ML-Algorithmen aus, um wichtige Designprobleme in Computersystemen anzugehen: die Energieeffizienz, die Leistung und die Chipfläche. Durch Quantisierung, bei der die Anzahl der zur Darstellung von Zahlen verwendeten Bits reduziert wird, können der Stromverbrauch und der Datenfluss erheblich reduziert und dadurch die Energieeffizienz verbessert werden, indem flächensparendere Chips in kompakten Recheneinheiten eingesetzt werden. Diese Techniken führen häufig zu einer Rechenbeschleunigung aufgrund reduzierter Datenpaketgrößen. Dadurch wird eine schnellere, parallelere und in Pipelines ausgeführte Verarbeitung gefördert, insbesondere bei der Berechnung neuronaler Netze. Andererseits führt dieser Ansatz jedoch zu einem Kompromiss zwischen Zahlengenauigkeit und Modellgenauigkeit, was geeignete Methoden für den Hardwareentwurf erfordert. Besonders im Hinblick auf Beschleuniger mit benutzerdefinierter Gleitkommaberechnung (FP) gibt es trotz der Fortschritte bei den Methoden des Stands der Technik immer noch erheblichen Raum für weiterführende Forschung.

In dieser Dissertation wird eine Hardware-Design-Methodik für Low-Power-Inferenz von neuronalen Spike-by-Spike (SbS)-Netzen für eingebettete Anwendungen im Bereich der Spiking Neural Networks (SNNs) vorgestellt. Im Vergleich zu herkömmlichen SNNs, die den Leaky Integrate-and-Fire (LIF)-Mechanismus verwenden, werden neuronale SbS-Netzwerke wegen ihrer reduzierten Modellkomplexität und außergewöhnlichen Rauschrobustheit beleuchtet. Trotz ihrer Vorteile haben SbS-Netzwerke jedoch von Natur aus einen Speicherplatzbedarf und Rechenkosten, die den Einsatz in eingeschränkten eingebetteten Systemen zu einer Herausforderung machen. Um dieses Problem zu lösen, verfolgt diese Forschungsarbeit die intrinsische Fehlerresilienz von SbS-Modellen zur Leis-

tungsverbesserung und Reduktion der Hardwarekomplexität bei gleichzeitiger Vermeidung von Zahlenquantisierung. Insbesondere führt diese Forschungsarbeit ein neuartiges Multiply-Accumulate (MAC)-Modul ein, das entwickelt wurde, um das Gleichgewicht zwischen Rechengenauigkeit und Ressourceneffizienz von FP-Operationen zu optimieren. Dieses MAC-Modul bietet konfigurierbare Qualität durch einen hybriden Ansatz. Es kombiniert Standard-FP-Zahlendarstellungen mit einem benutzerdefinierten 8-Bit-FP-Format sowie einer logarithmischen 4-Bit-Zahlendarstellung. Ferner kommt dieses Design ohne Verwendung eines Vorzeichenbits aus und trägt somit weiter zur kompakten und effizienten Darstellung von Zahlen bei. Darüber hinaus ermöglicht dieses Design, das MAC-Modul an die spezifischen Ressourcenbeschränkungen und Leistungsanforderungen einer bestimmten Anwendung anzupassen, wodurch neuronale SbS-Netzwerke für den Einsatz in Umgebungen mit eingeschränkten Ressourcen bereitgestellt werden können.

Im Bereich der Convolutional Neural Networks (CNNs) stellt diese Dissertation eine Hardware-Design-Methodik für Low-Power-Inferenz vor, die speziell auf Sensor-Analyse-Anwendungen abzielt. Im Mittelpunkt dieser Arbeit steht der Vorschlag für das Quantisierungsschema Hybrid-Float6 (HF6) und sein dedizierter Hardwarebeschleuniger, der als Conv2D-Tensorprozessor (TP) fungieren soll. Diese Quantisierungsstrategie verwendet eine hybride Zahlendarstellung, welche Standard-FP mit einem 6-Bit-FP-Format kombiniert. Diese Strategie ermöglicht einen hochoptimierten FP-MAC, der die Mantissenmultiplikation auf eine Multiplexer-Addierer-Operation reduziert. Diese Forschungsarbeit führt eine Quantization-Aware Training (QAT)-Methode ein, die in bestimmten Fällen vorteilhafte Regularisierungseffekte bietet. Die Wirksamkeit dieses Ansatzes wird in einem Regressionsmodell demonstriert, das trotz der angewendeten Quantisierung eine verbesserte Genauigkeit zeigt. Für die ML-Portabilität wird die benutzerdefinierte FP-Darstellung in ein Standardformat gekapselt - ein Designmerkmal, das es der vorgeschlagenen Hardware ermöglicht, sie automatisch zu verarbeiten. Um die Interoperabilität dieses Ansatzes zu validieren, wird die Hardware-Architektur in TensorFlow Lite integriert. Hiermit wird die Kompatibilität zum Industriestandard-ML-Frameworks demonstriert und das Potenzial für den praktischen Einsatz in verschiedenen Sensoranwendungen unter Beibehaltung der Einhaltung der etablierten ML-Infrastruktur bestätigt.

Diese Dissertation befasst sich mit einer wesentlichen Herausforderung in der aktuellen technologischen Landschaft: der Harmonisierung von Rechengenauigkeit mit Energieeffizienz und der Kompatibilität von Hardwarelösungen. Sie leistet einen wesentlichen Beitrag zur Entwicklung einer nachhaltigen nächsten Generation von neuronalen Netzwerkprozessoren, die für die Stärkung der zunehmend vernetzten und intelligenten Welt von morgen unerlässlich sind.