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

December 6, 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 Expansion von AI läutet eine neue Ära ein, die durch allgegenwärtige, vernetzte Geräte gekennzeichnet ist. Um die Nachhaltigkeit dieser Transformation zu gewährleisten, ist es unerlässlich, Entwurfsstrategien zu adoptieren, die präzise Rechenergebnisse mit wirtschaftlich tragfähigen Systemarchitekturen in Einklang bringen. Folglich steht die Verfeinerung der Effizienz und Qualität von AI-Hardwaremotoren als kritische Überlegung in dieser Entwicklung. Dies erfordert einen ausgewogenen Ansatz, der energieeffiziente Berechnungen, präzise und zuverlässige Ergebnisse sowie die Integration über verschiedene Plattformen und Geräte priorisiert.

ML-Algorithmen dienen als grundlegender Ermöglicher für die Integration von AI in IoT-Geräte, insbesondere im Kontext von Industrie 4.0. Diese Fortschritte prägen Anwendungen, um intelligenter und wirtschaftlich lohnender zu werden. Diese Transformation verbessert zahlreiche Bereiche, von der wissenschaftlichen Forschung über industrielle Prozesse bis hin zum alltäglichen Leben. Diese technologische Entwicklung bringt jedoch ihre eigenen Herausforderungen mit sich. ML-Algorithmen stellen erhebliche rechnerische und energetische Anforderungen. Folglich ist ein zentrales Ziel dieser Dissertation, innovative Methoden zur Steigerung der Hardwareeffizienz von Rechenmotoren zu erforschen.

Approximative Rechentechniken, wie Quantisierung, nutzen die inhärente Fehlerresilienz von ML-Algorithmen, um Schlüsseldesignanliegen in Computersystemen anzugehen: Energieeffizienz, Leistung und Chipfläche. Quantisierung, die die Reduzierung der zur Repräsentation von Zahlen verwendeten Bitanzahl beinhaltet, kann den Energieverbrauch und die Datenbewegung erheblich senken und somit die Energieeffizienz durch den Einsatz kompakter arithmetischer Einheiten verbessern, die Chipfläche sparen. Diese Techniken führen oft zu einer Beschleunigung der Berechnung aufgrund reduzierter Datengrößen, was schnellere, parallele und gepipelte Verarbeitung fördert, insbesondere in der Berechnung neuronaler Netzwerke. Dieser Ansatz führt jedoch zu einem Kompromiss zwischen Präzision und Modellgenauigkeit, der eine angemessene Hardware-Designmethodik erfordert. Während modernste Methoden voranschreiten, bleiben bedeutende Forschungsmöglichkeiten bestehen, insbesondere für Beschleuniger mit benutzerdefinierter FP-Berechnung.

In dieser Dissertation wird eine Hardware-Designmethodik für stromsparende Inferenz von SBS-neuronalen Netzwerken für eingebettete Anwendungen im Bereich der SNNs vorgestellt. Im Vergleich zu konventionellen SNNs, die den LIF-Mechanismus verwenden, werden SBS-neuronale Netzwerke für ihre reduzierte Modellkomplexität und außergewöhnliche Geräuschrobustheit hervorgehoben. Trotz ihrer Vorteile besitzen SBS-Netzwerke jedoch einen inhärenten Speicherbedarf und Rechenkosten, die ihre Bereitstellung in ressourcenbeschränkten Geräten herausfordernd machen. Um dieses Problem zu lösen, nutzt diese Forschung die inhärente Fehlerresilienz von SBS-Modellen, um die Leistung zu steigern und die Hardwarekomplexität zu reduzieren, während Quantisierung vermieden wird. Insbesondere führt diese Forschung ein neuartiges MAC-Modul ein, das darauf ausgelegt ist, das Gleichgewicht zwischen Rechengenauigkeit und Ressourceneffizienz von FP-Operationen zu optimieren. Dieses MAC-Modul verfügt über eine konfig-

urierbare Qualität durch einen hybriden Ansatz. Es kombiniert standardmäßige FP-Zahldarstellungen mit einem benutzerdefinierten 8-Bit-FP-Format sowie einer 4-Bit-logarithmischen Zahldarstellung. Dieses Design schließt die Verwendung eines Vorzeichenbits aus und trägt weiter zur kompakten und effizienten Darstellung von Zahlen bei. Dieses Design ermöglicht es, das MAC-Modul an die spezifischen Ressourcenbeschränkungen und Leistungsanforderungen einer bestimmten Anwendung anzupassen, und macht SBS-neuronale Netzwerke für den Einsatz in ressourcenbeschränkten Umgebungen möglich.

Im Bereich der CNNs präsentiert diese Dissertation eine Hardware-Designmethodik für stromsparende Inferenz, die speziell auf Sensoranalytikanwendungen abzielt. Zentral für diese Arbeit ist der Vorschlag des HF6-Quantisierungsschemas und seines dedizierten Hardware-Beschleunigers, der als Conv2D TP fungiert. Diese Quantisierungsstrategie verwendet eine hybride Zahldarstellung, die standardmäßige FP und ein 6-Bit-FP-Format kombiniert. Diese Strategie ermöglicht ein hochgradig optimiertes FP MAC, indem die Mantissenmultiplikation in eine Multiplexer-Addierer-Operation reduziert wird. Diese Forschung führt eine QAT-Methode ein, die in bestimmten Fällen vorteilhafte Regularisierungseffekte bietet. Die Wirksamkeit dieser Untersuchung wird mit einem Regressionsmodell demonstriert, das seine Präzision trotz der angewandten Quantisierung verbessert. Für die Portabilität von ML wird die benutzerdefinierte FP-Darstellung in einem Standardformat eingekapselt - eine Designeigenschaft, die es der vorgeschlagenen Hardware ermöglicht, sie automatisch zu verarbeiten. Um die Interoperabilität dieses Ansatzes zu validieren, wird die Hardware-Architektur in TensorFlow Lite integriert, was die Kompatibilität mit branchenüblichen ML-Frameworks demonstriert und das Potenzial für praktische Einsätze in verschiedenen Sensoranwendungen unter Beibehaltung der Übereinstimmung mit etablierter ML-Infrastruktur bestätigt.

Diese Dissertation behandelt eine wesentliche Herausforderung in der aktuellen technologischen Landschaft: die Harmonisierung von Rechengenauigkeit mit Energieeffizienz und der Kompatibilität von Hardwarelösungen. Diese Dissertation stellt einen bedeutenden Beitrag zur Entwicklung einer nachhaltigen nächsten Generation von neuronalen Netzwerkprozessoren dar, die für die zunehmend vernetzte und intelligente Welt von morgen unerlässlich sind.