Energiespar-Beschleunigungschips für neuronale Netzwerke: Fortschritte bei benutzerdefinierten Gleitkommatechniken.

Zusammenfassung

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 Leistungsverbesserung 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.