Patent-Exposé

Brandenburgische Technische Universität Cottbus - Senftenberg

Vorrichtung und Verfahren zur Berechnung eines Matrix-Vektor-Produkts einer bestimmten Matrix und eines Vektors

Angebot Nr.	22-01
Stand	September 2026

Hintergrund

Künstliche neuronale Netze gewinnen in Bereichen wie der Bild- und Sprachverarbeitung massiv an Bedeutung, verursachen jedoch bei der Berechnung komplexer Matrix-Vektor-Multiplikationen einen enormen und energieintensiven Rechenaufwand. Um die Effizienz von Hardwarebeschleunigern (wie GPUs oder TPUs) zu steigern, nutzen bisherige Ansätze Optimierungen wie die Quantisierung von Gewichten oder das Pruning, welches durch das gezielte Nullsetzen von Werten zu dünnbesetzten Matrizen führt. Dennoch stoßen diese Verfahren bei der Reduzierung der Rechenlast an technologische Grenzen oder erfordern eine hochkomplexe Steuerungslogik. Zudem mangelt es herkömmlichen Algorithmen an der nötigen Flexibilität, um Berechnungen auf rekonfigurierbarer Hardware ohne signifikante Einbußen bei der Genauigkeit effizient und ressourcenschonend abzubilden.

Beschreibung

Die Erfindung beschreibt eine Vorrichtung und ein computerimplementiertes Verfahren zur ressourceneffizienten Berechnung eines Matrix-Vektor-Produkts, das insbesondere für die rechenintensiven Inferenzprozesse in künstlichen neuronalen Netzen (ANNs) eingesetzt wird. Das Kernprinzip des Verfahrens basiert auf der mathematischen Zerlegung (faktoriellen Dekomposition) einer gegebenen Gewichtsmatrix in vertikale Abschnitte beziehungsweise Teilmatrizen, welche wiederum als Produkt mehrerer hintereinandergeschalteter, extrem dünnbesetzter Matrizen (sogenannter Sparse- oder CC-Matrizen) angenähert werden. Die Einträge dieser CC-Matrizen weisen eine vordefinierte Struktur auf und bestehen im Wesentlichen aus Nullen sowie vorzeichenbehafteten Zweierpotenzen.

Die zugehörige Hardware-Architektur bildet diese mathematische Struktur direkt ab, indem sie für jeden vertikalen Matrixabschnitt eine parallele Verarbeitungskette mit einer Kaskade seriell gekoppelter Verarbeitungsblöcke nutzt. Da die Multiplikation eines Vektors mit einer Zweierpotenz informationstechnisch rein durch eine Bitverschiebung dargestellt werden kann, verarbeiten die Blöcke die Eingangsvektoren ausschließlich über Shifter-Module und nachfolgende Addierer-Einheiten. Auf diese Weise wird die rechen- und energieintensive Nutzung klassischer Hardware-Multiplizierer vollständig eliminiert. Je nach Anwendungsfall kann das Schaltungsdesign als feste Verdrahtung in integrierten Schaltkreisen (ASICs) realisiert oder zur flexiblen Anpassung an wechselnde Netzwerk-Gewichte dynamisch auf programmierbarer Hardware (FPGAs) implementiert werden.

Referat Patente und Lizenzen
Zentralcampus Cottbus: +49 (0) 355 69 3535
www.b-tu.de/transfer/patente-lizenzen
patente@b-tu.de

Details / Vorteile

Vollständiger Verzicht auf Multiplizierer
Massive Einsparung von Hardware-Ressourcen
Frei skalierbare Berechnungsgenauigkeit
Optimierte Pipeline-Architektur
Vielseitige Hardware-Kompatibilität

Anwendungsfälle

Edge-AI und mobile Endgeräte
Fahrerassistenzsysteme (ADAS) & Automotive
Automatisierte Bild- und Sprachverarbeitung
Echtzeit-Signalverarbeitung & Kommunikation
Massen-Empfehlungssysteme (DLRM) in Rechenzentren

Schutzrecht

EP4318276

Entwicklungsstand

Technologiekonzept

Angebote

- Ausgründung
- FuE-Kooperation
- Lizenzierung

Kontakt

Mike König
Platz der Deutschen Einheit 1
03046 Cottbus
Hauptgebäude, Zentralcampus
Raum HG 2.39

T +49 (0) 355 69-3535
E patente@b-tu.de