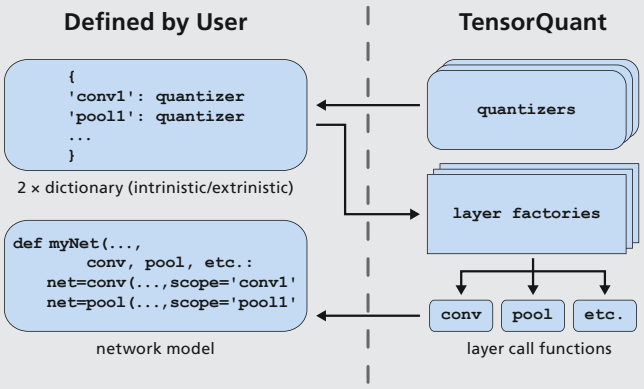
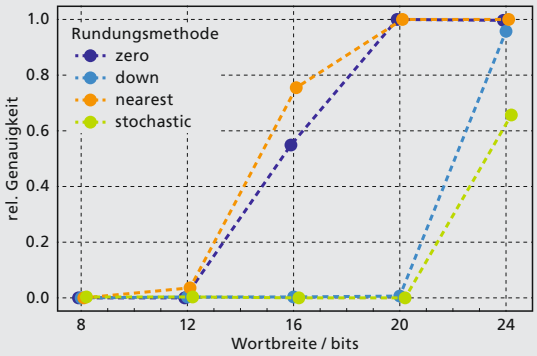


1



2

TENSORQUANT BRINGT DEEP LEARNING IN DIE MOBILE ANWENDUNG

Der Trend, Methoden des Maschinellen Lernens, insbesondere tiefer künstlicher neuronaler Netze (Deep Learning – DL), zur Entwicklung intelligenter Systeme einzusetzen, hat sich weiter verstärkt. Aus der Wissenschaft kommen dabei stetig neue Deep-Learning-Ansätze, welche weit gefächerte Einsatzmöglichkeiten dieser Algorithmen in vielen verschiedenen praktischen Anwendungen ermöglichen könnten.

Eine entscheidende technologische Hürde auf dem Weg hin zur Anwendung in Massenprodukten ist jedoch der enorme Rechenaufwand, der zur Berechnung und Auswertung der DL-Modelle notwendig ist.

Daher ist in der letzten Zeit die Entwicklung spezialisierter DL-Hardware in den Fokus getreten. Neue Chip- und Speicherarchitekturen sollen zukünftig den Einsatz performanter und zugleich energiesparender Hardwaremodule ermöglichen und so den Einsatz von DL in z. B. autonomen Fahrzeugen, Mobiltelefonen oder integrierten Produktionssteuerungen ermöglichen.

Lernen erfordert keine hohe Präzision in der Zahldarstellung

Dabei nutzt man eine wesentliche mathematische Eigenschaft des DL aus: Das Lernen und Auswerten von Modellen lässt sich auf die numerische Berechnung einer recht kleinen Anzahl von Operationen aus der Tensor-Algebra (wie z. B. Matrixmultiplikationen) reduzieren. Außerdem kommt man bei der Tensor-Berechnung mit einer deutlich geringeren Genauigkeit in der Zahldarstellung aus, als man dies typischerweise von physikalischen Simulationen gewohnt ist. Diese Eigenschaften ermöglichen nun eine – im Vergleich zu allgemeinen Recheneinheiten wie CPUs und GPUs – hocheffiziente Hardwareumsetzung.

TensorQuant erlaubt die Simulation von Machine Learning Hardware

Bei der Entwicklung von DL-Anwendungen auf spezialisierter Hardware ergibt sich allerdings die Schwierigkeit, dass die Mindestanforderungen an die Rechengenauigkeit zwischen einzelnen Modellen stark variieren. Dies macht die gleichzeitige Optimierung von DL-Modellen und Hardware bezüglich Rechenperformanz, Energieverbrauch und Vorhersagegenauigkeit schwierig. Mit unserem Softwaretool TensorQuant (TQ) können Entwickler nun DL-Modelle mit beliebigen Zahldarstellungen und Rechengenauigkeiten simulieren, kritische Tensor-Operationen identifizieren und damit diesen Entwicklungsschritt deutlich beschleunigen. TQ wird bereits in Kooperationsprojekten mit der Automobilindustrie eingesetzt.

1 *TensorQuant erlaubt die automatische Simulation von gegebenen TensorFlow-Modellen mit beliebigen Zahldarstellungen einzelner Tensor-Operationen.*

2 *Das Ergebnis einer Simulation des bekannten ResNet-50 Models zeigt, dass die konkrete Wahl der Zahldarstellung erheblichen Einfluss auf die Performanz von DL-Anwendungen hat, welche ohne die Simulation in TensorQuant vorab nur schwer abzuschätzen ist.*