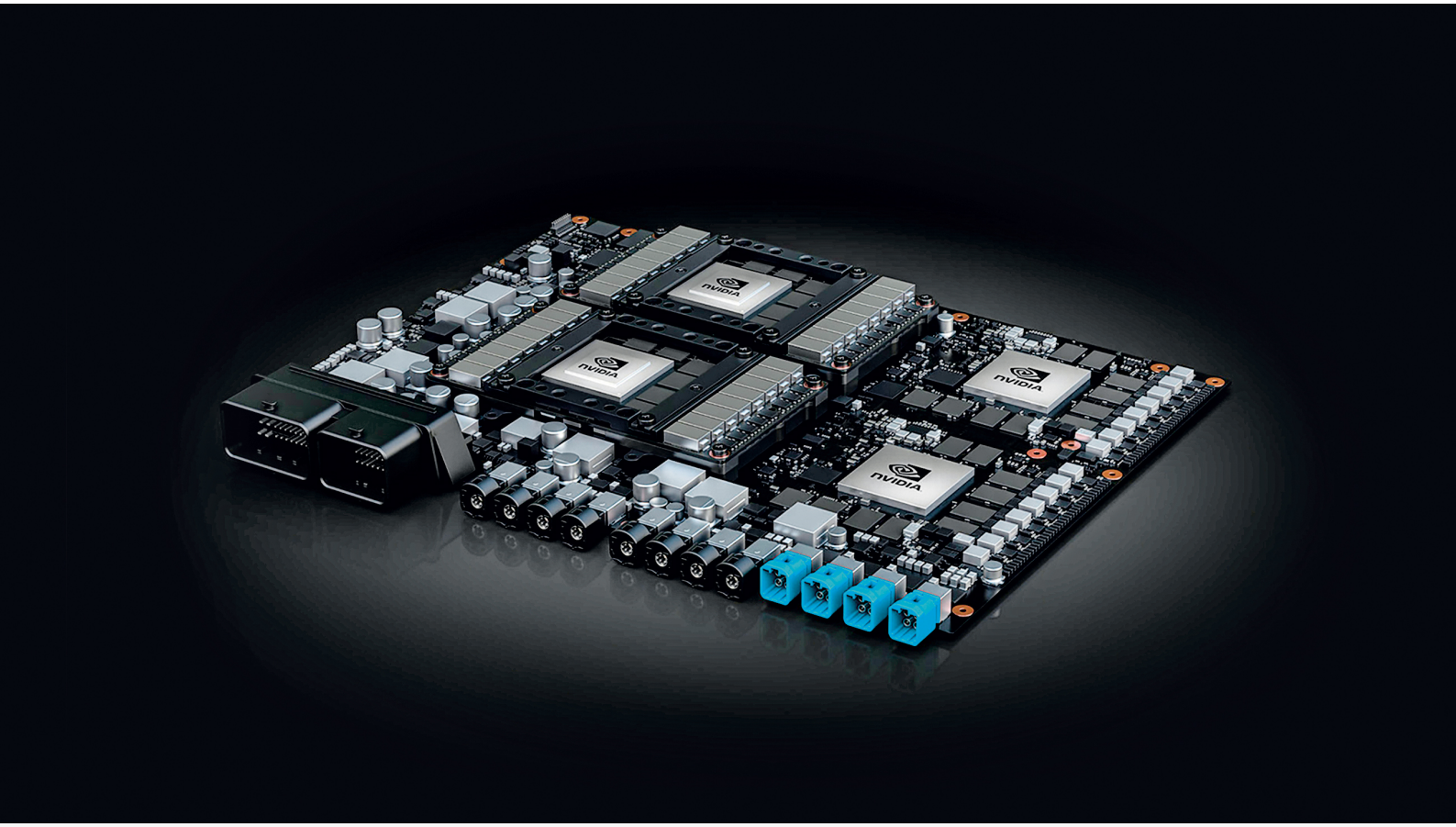


Booster für KI-Berechnungen

Dieser Artikel gibt den Stand zum Zeitpunkt seiner Veröffentlichung wieder. Der Inhalt kann inzwischen veraltet sein.

13/02/2020 Automatisierte und autonome Fahrfunktionen lassen sich ohne KI nicht realisieren. Die dafür benötigte Rechenleistung liefern Spezial-Chips, die auf paralleles Rechnen spezialisiert sind. Forscher arbeiten aber auch an neuen biologisch inspirierten Lösungen sowie an Quantencomputern, die noch mehr Rechenleistung versprechen.

Seit Jahrzehnten hält immer mehr Elektronik Einzug ins Fahrzeug. Heute kontrollieren Dutzende vernetzte Steuergeräte Motor, Getriebe, Infotainmentsystem und viele weitere Funktionen. Autos sind also längst rollende Rechenzentren – aber jetzt steht ihnen ein neuer Sprung in puncto Computerpower bevor, denn automatisierte Fahrfunktionen und das autonome Fahren erfordern noch leistungsstärkere Rechner. Und weil die nötige Performance mit herkömmlichen Chips nicht mehr zu erreichen ist, schlägt jetzt die Stunde von Grafikprozessoren, Tensor Processing Units (TPUs) und anderer Hardware, die speziell für Berechnungen von neuronalen Netzen ausgelegt ist.

Klassische CPUs (Central Processing Units) sind zwar universell einsetzbar, verfügen für KI aber nicht

über die optimale Architektur. Das liegt an den typischen Berechnungen, die beim Training von und beim Schlussfolgern (Inferenz) mit neuronalen Netzen auftreten. „Die Matrizenmultiplikationen in neuronalen Netzen sind sehr aufwendig", erklärt Dr. Markus Götz vom Steinbuch Centre for Computing am Karlsruher Institut für Technologie (KIT). „Aber die Berechnungen lassen sich wunderbar parallelisieren – besonders mit Grafikkarten. Schafft eine Highend-CPU mit 24 Kernen und Vektorbefehlen 24 mal 4 Berechnungen pro Zyklus, so sind es bei einer modernen Grafikkarte über 5.000."

Grafikprozessoren (GPUs, Graphic Processing Units) sind von vornherein auf paralleles Arbeiten spezialisiert und verfügen dafür über eine maßgeschneiderte interne Architektur: GPUs enthalten Hunderte oder Tausende einfache Rechenkerne für Ganzzahl- und Gleitkommaoperationen, die simultan die gleiche Operation auf verschiedene Daten anwenden können (Single Instruction Multiple Data). So sind sie in der Lage, Tausende Rechenoperationen pro Taktzyklus durchzuführen – etwa um die Pixel einer virtuellen Landschaft zu berechnen oder die Matrizenmultiplikationen für neuronale Netze durchzuführen. Kein Wunder also, dass Chips des GPU-Herstellers NVIDIA derzeit als Arbeitspferde für Künstliche Intelligenz im Allgemeinen und fürs autonome Fahren im Besonderen hoch im Kurs stehen. Unter anderem setzt Volkswagen auf die Hardware des US-Unternehmens. „Man braucht spezielle Hardware für das autonome Fahren", sagt Ralf Bauer, Leiter Software-Entwicklung bei Porsche Engineering. „GPUs sind der Anfang, später werden wahrscheinlich anwendungsspezifische Chips folgen."

Aktuell bietet NVIDIA den Prozessor Xavier speziell für das autonome Fahren an. Auf einem Silizium-Chip sind unter anderem acht herkömmliche CPUs und eine speziell für das maschinelle Lernen optimierte GPU untergebracht. Für automatisiertes Fahren auf Stufe 2+ (eingeschränkte Längs- und Querführung mit im Vergleich zu Stufe 2 erweiterter Funktionalität auf Basis der Standard-Sensoren) steht die Drive AGX Xavier-Plattform zur Verfügung, die maximal 30 Billionen Rechenoperationen pro Sekunde durchführen kann (30 TOPS, Tera Operations Per Second). Für hochautomatisiertes und autonomes Fahren hat NVIDIA den KI-Computer Drive AGX Pegasus (320 TOPS) im Programm, unter dessen Kontrolle ein Versuchsfahrzeug bereits 80 Kilometer ohne menschlichen Eingriff durchs Silicon Valley gefahren ist. Als Nachfolger des Xavier entwickelt NVIDIA derzeit die GPU Orin, über deren Leistungsdaten bisher nur wenig bekannt ist.

Aber nicht alle Fahrzeughersteller setzen auf GPUs. 2016 hat Tesla damit begonnen, eigene Prozessoren für neuronale Netze zu entwickeln. Statt der Grafikprozessoren von NVIDIA baut das US-Unternehmen seit Frühjahr 2019 seinen FSD-Chip (Full Self Driving) in seine Fahrzeuge ein. Er enthält neben zwei „Neural Processing Units" (NPUs) mit jeweils 72 TOPS auch zwölf herkömmliche CPU-Kerne für allgemeine Berechnungen und eine GPU für das Postprocessing von Bild- und Videodaten. Die NPUs sind – ähnlich wie GPUs – auf die parallele und damit schnelle Ausführung von Additionen und Multiplikationen spezialisiert.

Google-Chip für KI-Anwendungen

Google ist ein weiterer Newcomer im Chip-Geschäft: Das Technologieunternehmen nutzt seit 2015 die selbst entwickelten TPUs in seinen Rechenzentren. Der Name leitet sich vom mathematischen Begriff „Tensor" ab, unter den unter anderem Vektoren und Matrizen fallen. Darum heißt Googles weitverbreitete Software-Bibliothek für Künstliche Intelligenz auch „TensorFlow" – und für sie sind die Chips auch optimiert. 2018 hat Google die dritte Generation seiner TPUs vorgestellt, die vier „Matrizenmultiplikationseinheiten" enthalten und insgesamt 90 TFLOPS (Tera Floating Point Operations Per Second) erreichen sollen. Die Google-Tochter Waymo nutzt TPUs, um neuronale Netzwerke für das autonome Fahren zu trainieren.

Anwendungsspezifische Chips wie Teslas FSD oder die TPUs von Google rechnen sich erst ab großen Stückzahlen. Eine Alternative sind FPGAs (Field Programmable Gate Arrays): Diese universell einsetzbaren digitalen Chips enthalten zahlreiche Rechen- und Speicherblöcke, die sich per Programmierung miteinander kombinieren lassen und mit denen man Algorithmen quasi in Hardware gießen kann – wie bei einem anwendungsspezifischen Chip, aber zu wesentlich geringeren Kosten. So lassen sich FPGAs einfach an die spezifischen Anforderungen einer KI-Anwendung anpassen (etwa an vorgegebene Datentypen), was zu Vorteilen bei Performance und Energieverbrauch führt. Das Münchener Start-up Kortiq hat für FPGAs die „AIScale"-Architektur entwickelt, die neuronale Netzwerke für die Bilderkennung so vereinfacht und die Berechnungen so optimiert, dass die Anforderungen an die Hardware deutlich sinken und die Ergebnisse bis zu zehnmal schneller zur Verfügung stehen.

Manche Forscher setzen für KI-spezifische Chips auf eine noch engere Anlehnung an die Arbeitsweise von Nervenzellen. An der Universität Heidelberg ist das neuromorphe System „BrainScaleS" entstanden, dessen künstliche Neuronen als analoge Schaltungen in Silizium-Chips realisiert sind: Der Zellkörper besteht aus rund 1.000 Transistoren und zwei Kondensatoren, die Synapsen benötigen etwa 150 Transistoren. Einzelne Zellkörper lassen sich wie in einem Baukasten zu verschiedenen Typen künstlicher Neuronen kombinieren. Die Synapsen können wie in der Natur unterschiedlich starke Verbindungen aufbauen, außerdem gibt es erregende und hemmende Typen. Der Output der Neuronen besteht aus „Spikes" – kurzen Spannungsimpulsen von wenigen Mikrosekunden Dauer, die den anderen künstlichen Neuronen als Inputs dienen.

Energieeffiziente Neuro-Chips

BrainScaleS dient aber nicht nur der Erforschung des menschlichen Gehirns. Mit den künstlichen Neuronen ließen sich auch technische Probleme lösen – etwa die Objekterkennung fürs autonome Fahren. Denn sie bieten einerseits eine hohe Rechenleistung von etwa einer Billiarde Rechenoperationen (1.000 TOPS) pro Modul mit 200.000 Neuronen. Andererseits verbraucht die analoge Lösung auch sehr wenig Energie. „Bei digitalen Schaltungen sind zum Beispiel für jede Operation etwa 10.000 Transistoren im Einsatz", erklärt Johannes Schemmel von der Universität Heidelberg. „Wir kommen mit

wesentlich weniger aus, wodurch wir ungefähr 100 TOPS pro Watt erreichen können." Die Forscher haben gerade die zweite Generation ihrer Schaltungen entwickelt und sprechen mit Industriepartnern über mögliche Kooperationen.

Quanten-Power aus der Cloud

In Zukunft könnten auch Quantencomputer im Bereich KI zum Einsatz kommen. Ihre fundamentale Einheit ist nicht das zweiwertige Bit, sondern das Qubit mit unendlich vielen möglichen Werten. Dank der Gesetze der Quantenmechanik lassen sich Berechnungen stark parallelisieren und damit beschleunigen. Allerdings sind Quantencomputer nur schwer zu realisieren, weil die Qubits durch empfindliche physikalische Systeme wie Elektronen, Photonen oder Ionen repräsentiert werden. Das zeigt sich zum Beispiel beim „IBM Q System One", den das Unternehmen auf der Elektronikmesse CES 2019 in Las Vegas vorgestellt hat: Das Innere des Quantencomputers muss penibel von Erschütterungen, elektrischen Feldern und Temperaturschwankungen abgeschirmt werden.

Kaufen kann man den IBM-Rechner nicht, er lässt sich aber über eine Cloud nutzen. Andere Hersteller wie D-Wave aus Kanada bieten ebenfalls Quantenpower an. Volkswagen hat zum Beispiel mit einem Quantencomputer des Unternehmens ein Verkehrsmanagementsystem realisiert, das die Effizienz von Taxiunternehmen und anderen Transportanbietern in urbanen Räumen verbessern soll. „Neuronale Netze sind auch eine Art Optimierungsaufgabe – in der Regel will man eine optimale Vorhersagegenauigkeit erreichen", sagt KIT-Experte Götz. „Darum wird derzeit unter anderem daran geforscht, wie man KI-Algorithmen für Quantencomputer ändern müsste."

Nervenzellen und künstliche Neuronen

Nervenzellen erhalten ihre Signale von anderen Neuronen über Synapsen, die sich entweder an den Dendriten oder direkt am Zellkörper befinden. Synapsen können entweder erregend oder hemmend wirken. Alle Inputs werden am Axonhügel summiert, und wenn dabei eine Schwelle überschritten wird, feuert die Nervenzelle ein etwa eine Millisekunde langes Signal ab, das sich längs des Axons fortpflanzt und andere Neuronen erreicht.

Künstliche Neuronen ahmen dieses Verhalten mehr oder weniger genau nach. In herkömmlichen neuronalen Netzen mit mehreren Schichten erhält jede „Nervenzelle" eine gewichtete Summe als Input. Sie besteht aus den Outputs ai der Neuronen in der vorgelagerten Schicht und den Gewichtsfaktoren wi , in denen die Lernerfahrung des neuronalen Netzes gespeichert ist. Diese Gewichtsfaktoren entsprechen den Synapsen und können ebenfalls erregend oder hemmend wirken. Ein einstellbarer Schwellwert bestimmt ähnlich wie bei der Nervenzelle, wann das künstliche Neuron feuert.

Lernen von und Schließen mit neuronalen Netzen

Natürliche und künstliche neuronale Netzwerke lernen durch Veränderungen der Stärke der synaptischen Verbindungen bzw. der Gewichtsfaktoren. In tiefen neuronalen Netzwerken legt man beim Training Daten an ihre Eingänge und vergleicht den Output mit einem gewünschten Ergebnis. Mithilfe mathematischer Verfahren werden die Gewichtsfaktoren wij so lange angepasst, bis das neuronale Netz beispielsweise Bilder zuverlässig in vorgegebene Kategorien einordnet. Beim Schließen legt man Daten an den Eingang und nutzt den Output zum Beispiel für Entscheidungen.

Sowohl beim Training als auch beim Schließen in tiefen neuronalen Netzwerken (Netzwerke mit mehreren Schichten von künstlichen Neuronen) treten immer wieder die gleichen mathematischen Operationen auf. Fasst man die Ausgänge der Neuronen in Schicht 1 und die Eingänge der Neuronen in Schicht 2 jeweils als Spaltenvektoren zusammen, lassen sich alle Berechnungen als Matrizenmultiplikationen darstellen. Dabei treten zahlreiche, voneinander unabhängige Multiplikationen und Additionen auf, die sich parallel ausführen lassen. Dafür sind herkömmliche CPUs nicht ausgelegt – und darum sind ihnen Grafikprozessoren, TPUs und andere KI- Beschleuniger weit überlegen.

Zusammengefasst

Bei Berechnungen für neuronale Netze stoßen klassische Computerchips an ihre Grenzen. Weit leistungsfähiger sind Grafikprozessoren sowie Spezial-Hardware für KI, die von Unternehmen wie NVIDIA und Google entwickelt werden. Neuromorphe Chips orientieren sich stark an echten Neuronen und arbeiten sehr energieeffizient. Quantencomputer könnten die Rechenleistungen nochmals enorm steigern.

Info

Text: Christian Buck
Mitwirkende: Ralf Bauer, Dr. Christian Koelen

Text erstmalig erschienen im Porsche Engineering Magazin, Nr. 2/2019

Linksammlung

Link zu diesem Artikel
https://newsroom.porsche.com/de/2020/technik/porsche-engineering-booster-ki-berechnungen-autonomes-fahren-19488.html

Externe Links
https://www.porscheengineering.com/peg/de/