Auf dem Quantensprung

22/04/2022 Immer mehr Fahrzeugfunktionen basieren auf Künstlicher Intelligenz. Herkömmliche Prozessoren und Grafikchips stoßen bei den Berechnungen Neuronaler Netze zunehmend an ihre Grenzen. Porsche Engineering berichtet über neue Technologien, die KI-Berechnungen in Zukunft beschleunigen sollen.

Kunstliche Intelligenz (KI) ist eine Schlusseltechnologie fur die Automobilbranche – entsprechend wichtig ist schnelle Hardware fur die aufwendigen Backend-Berechnungen, die damit verbunden sind. Denn nur mit Hochleistungscomputern lassen sich neue Funktionen kunftig in Serie bringen. „Autonomes Fahren gehort zu den anspruchsvollsten KI-Anwendungen uberhaupt“, erklart Dr. Joachim Schaper, Leiter Fachdisziplin KI und Big Data bei Porsche Engineering. „Die Algorithmen lernen anhand sehr vieler Beispiele, die Testfahrzeuge zuvor per Kamera, Radar oder anderen Sensoren im realen Verkehr gesammelt haben.“

Herkommliche Rechenzentren sind den wachsenden Anforderungen immer weniger gewachsen. „Inzwischen dauert es Tage, um eine einzige Variante eines Neuronalen Netzes zu trainieren“, erklart Schaper. Fur ihn steht darum fest: Die Automobilhersteller brauchen neue Technologien fur KI-Berechnungen, mit deren Hilfe die Algorithmen sehr viel schneller lernen konnen. Um das zu erreichen,

mussen moglichst viele Vektor-Matrix-Multiplikationen in den komplexen Neuronalen Netzen (englisch Deep Neural Networks, kurz DNN) parallel ausgefuhrt werden – eine Aufgabe, auf die Grafikprozessoren (GPUs) spezialisiert sind. Ohne sie waren die erstaunlichen Fortschritte der KI in den letzten Jahren nicht moglich gewesen.

50-mal so groß wie ein Grafikprozessor

Allerdings wurden Grafikkarten ursprunglich nicht fur den KI-Einsatz entworfen, sondern um Bilddaten moglichst effizient zu verarbeiten. Sie stoßen darum beispielsweise beim Training von Algorithmen fur das autonome Fahren zunehmend an ihre Grenzen. Fur noch schnellere Berechnungen ist deshalb auf KI spezialisierte Hardware erforderlich. Eine mogliche Losung hat die kalifornische Firma Cerebras prasentiert. Ihre „Wafer Scale Engine" (WSE) ist optimal auf die Anforderungen Neuronaler Netze zugeschnitten, indem sie moglichst viel Rechenkraft auf einem riesigen Computerchip vereint. Er ist mehr als 50-mal so groß wie ein normaler Grafikprozessor und bietet Platz fur 850.000 Rechenkerne – uber 100-mal so viele wie auf einem aktuellen Top-GPU.

Außerdem haben die Cerebras-Ingenieure die Rechenkerne untereinander mit Datenleitungen hoher Bandbreite vernetzt: Laut Hersteller transportiert das Netzwerk auf der Wafer Scale Engine 220 Petabit pro Sekunde. Auch den Flaschenhals innerhalb der GPUs hat Cerebras aufgeweitet: Zwischen Arbeitsspeicher und Rechenwerk reisen die Daten fast 10.000-mal schneller als in leistungsstarken GPUs – mit 20 Petabyte pro Sekunde.

Um noch mehr Zeit zu sparen, imitiert Cerebras einen Trick des Gehirns. Dort arbeiten Neuronen nur, wenn sie von anderen Neuronen Signale bekommen. Die vielen gerade inaktiven Verbindungen brauchen keine Ressourcen. In DNNs hingegen kommt es haufig vor, dass bei Vektor-Matrix-Multiplikation mit der Zahl Null multipliziert wird. Das kostet unnotig Zeit. Die Wafer Scale Engine unterlasst es deshalb. „Alle Nullen werden herausgefiltert", schreibt Cerebras in seinem Whitepaper zur WSE. Der Chip fuhrt also nur Operationen aus, die ein von null verschiedenes Ergebnis liefern.

Ein Nachteil des Chips ist sein hoher Bedarf an elektrischer Leistung von 23 kW. Das macht eine Wasserkuhlung erforderlich. Cerebras hat darum ein eigenes Server-Gehause fur den Einsatz in Rechenzentren entwickelt. In den Datenzentren einiger Forschungsinstitute wird die Wafer Scale Engine bereits getestet. Der Riesenchip aus Kalifornien konnte auch die Automobilentwicklung beschleunigen, glaubt Experte Schaper. „Mit ihm ließe sich theoretisch eine Woche Training auf nur wenige Stunden reduzieren", schatzt er. „Das muss die Technik in Praxistests jedoch erst noch unter Beweis stellen."

Licht statt Elektronen

So ungewohnlich der neue Chip auch ist: Wie seine konventionellen Vorganger arbeitet auch er mit herkommlichen Transistoren. Unternehmen wie Lightelligence und Lightmatter aus Boston wollen statt

der vergleichsweise langsamen Elektronik das viel schnellere Licht fur KI-Berechnungen nutzen und bauen dafur optische Chips. DNNs konnten damit „mindestens einige hundert Mal schneller arbeiten als elektronische“, schreiben Entwickler von Lightelligence.

Dafur nutzen Lightelligence und Lightmatter das Phanomen der Interferenz. Wenn sich Lichtwellen gegenseitig verstarken oder ausloschen, bilden sie ein Hell-Dunkel-Muster. Lenkt man die Interferenz auf eine bestimmte Weise, entspricht das neue Muster der Vektor-Matrix-Multiplikation des alten Musters. Die Lichtwellen konnen also „rechnen“. Um das praktisch umzusetzen, haben die Bostoner Entwickler winzige Lichtleiter in einen Silizium-Chip geatzt. Wie in einem Textilgewebe uberkreuzen sie sich mehrfach. An den Kreuzungen findet die Interferenz statt. Dazwischen regeln winzige Heizelemente den Brechungsindex des Lichtleiters, wodurch sich die Lichtwellen gegeneinander verschieben lassen. So kann man deren Interferenz steuern und Vektor-Matrix-Multiplikationen ausfuhren.

Ganz auf Elektronik verzichten die Bostoner Unternehmen aber nicht. Sie kombinieren ihre Lichtrechner mit herkommlichen elektronischen Bauelementen, die Daten speichern und alle Berechnungen außer den Vektor-Matrix-Multiplikationen ausfuhren. Dazu gehoren zum Beispiel die nichtlinearen Aktivierungsfunktionen, mit denen die Ausgabewerte jedes Neurons modifiziert werden, bevor sie in die nachste Schicht gelangen.

Mit der Kombination aus optischen und digitalen Rechnern lassen sich DNNs extrem schnell berechnen. „Ihr Hauptvorteil liegt in der geringen Latenzzeit“, erklart Lindsey Hunt, Sprecherin von Lightelligence. Das DNN kann dadurch beispielsweise schneller Objekte auf Bildern erkennen, wie etwa Fußganger und e-Scooter-Fahrer. Beim autonomen Fahren konnte dies zu schnelleren Reaktionen in kritischen Situationen fuhren. „Zudem trifft das optische System mehr Entscheidungen pro Watt elektrischer Energie“, so Hunt. Das ist besonders wichtig, weil die steigende Rechenleistung in Fahrzeugen zunehmend auf Kosten von Kraftstoffverbrauch und Reichweite geht.

Die Losungen von Lightmatter und Lightelligence lassen sich als Module in herkommliche Computer einsetzen, um die KI-Berechnungen zu beschleunigen – ahnlich wie Grafikkarten. Im Prinzip konnte man sie auch in Fahrzeuge integrieren, etwa um autonome Fahrfunktionen zu realisieren. „Unsere Technologie ist sehr gut geeignet, um als Inferenzmaschine fur ein autonomes Auto zu dienen“, erklart Lindsey Hunt. Experte Schaper sieht das ahnlich: „Wenn es Lightelligence gelingt, automobiltaugliche Komponenten zu bauen, konnte dies die Einfuhrung komplexer KI-Funktionen im Fahrzeug stark beschleunigen.“ Die Technik ist inzwischen marktreif: Das Unternehmen plant erste Pilotversuche mit Kunden fur das Jahr 2022.

Quantencomputer als KI-Turbo

Noch etwas weiter von der praktischen Anwendung entfernt sind Quantencomputer. Auch sie werden KI-Berechnungen beschleunigen, weil sie Unmengen an Daten parallel verarbeiten konnen. Sie arbeiten dafur mit sogenannten „Qubits“. Anders als die klassische Informationseinheit, das Bit, kann ein Qubit

die beiden binaren Werte 0 und 1 simultan darstellen. Die beiden Zahlen koexistieren in einem Uberlagerungszustand, wie er nur in der Quantenmechanik moglich ist.

Quantencomputer konnten beim Klassifizieren von Dingen zum Turbo fur Kunstliche Intelligenz werden, zum Beispiel im Verkehrsgeschehen. Dort gibt es viele unterschiedliche Objekt-Kategorien, darunter Fahrrader, Autos, Fußganger, Schilder, nasse und trockene Straßen. Sie unterscheiden sich anhand vieler Eigenschaften, weswegen Experten von „Mustererkennung in hoherdimensionalen Raumen" sprechen.

„Je komplizierter die Muster, desto schwerer tun sich herkommliche Rechner damit, Klassen zu unterscheiden", erklart Heike Riel, die IBMs Quantenforschung in Europa und Afrika leitet. Denn mit jeder Dimension wird es aufwendiger, die Ahnlichkeit zweier Objekte auszurechnen: Wie ahnlich sind sich ein e-Scooter-Fahrer und ein Rollatornutzer, die die Straße uberqueren wollen? Quantenrechner konnen im Vergleich zu klassischen Computern effizient in hochdimensionalen Raumen arbeiten. Fur gewisse Probleme konnte diese Eigenschaft nutzlich sein und dazu fuhren, dass mithilfe von Quantencomputern einige Probleme schneller gelost werden konnen als mit klassischen Hochleistungsrechnern.

IBM-Forscher haben statistische Modelle analysiert, die sich fur die Datenklassifizierung trainieren lassen. Erste Resultate deuten darauf hin, dass geschickt gewahlte Quantenmodelle fur gewisse Datensatze besser funktionieren als klassische Methoden. Die Quantenmodelle sind einfacher zu trainieren und scheinen eine großere Kapazitat zu haben – was ihnen erlaubt, kompliziertere Zusammenhange zu lernen.

Mit den heutigen Quantencomputern kann man diese Algorithmen testen, aber noch keinen Vorteil gegenuber klassischen Rechnern erzielen, räumt Riel ein. Die Entwicklung von Quantencomputern schreitet jedoch rasant voran. Sowohl die Anzahl der Qubits als auch deren Qualität nehmen stetig zu. Ein weiterer wichtiger Faktor ist die Geschwindigkeit, gemessen in Circuit Layer Operations per Second (CLOPS). Diese Zahl beschreibt, wie viele Quantenschaltkreise auf dem Quantenrechner pro Zeit laufen können. Sie ist eines der drei wichtigen Leistungskriterien eines Quantencomputers: Skalierbarkeit, Qualität und Geschwindigkeit.

In absehbarer Zeit durfte es gelingen, fur bestimmte Anwendungen die Uberlegenheit von Quantencomputern zu zeigen – also dass sie Probleme schneller, effizienter und praziser losen als ein klassischer Computer. Einen leistungsstarken, fehlerkorrigierten, universellen Quantenrechner zu bauen, dauert aber noch etwas langer. Nach Schatzung von Experten werden dafur noch mindestens zehn Jahre vergehen. Doch das Warten konnte sich lohnen. Wie auch optische Chips oder neue Architekturen von elektronischen Rechnern konnten Quantenrechner der Schlussel sein auf dem Weg zur Mobilitat der Zukunft.

Zusammengefasst

Bei KI-Berechnungen stoßen mittlerweile neben klassischen Mikroprozessoren auch Grafikchips an ihre Grenzen. Unternehmen und Forscher weltweit arbeiten darum an neuen Losungen. Nahe am Einsatz sind Chips im Wafer-Format und Lichtrechner. In einigen Jahren konnten Quantencomputer fur besonders anspruchsvolle Berechnungen hinzukommen.

Info

Text erstmals erschienen im Porsche Engineering Magazin, Ausgabe 1/2022.

Autor: Christian Meier

MEDIA ENQUIRIES

Inga Konen

Head of Communications Porsche Schweiz AG
+41 (0) 41 / 487 914 3
inga.konen@porsche.ch

Link Collection

Link to this article
https://newsroom.porsche.com/de_CH/2022/innovation/porsche-engineering-quantencomputer-computerchip-optische-rechner-28089.html

External Links
https://christophorus.porsche.com/en.html