



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 Schlüsseltechnologie für die Automobilbranche – entsprechend wichtig ist schnelle Hardware für die aufwendigen Backend-Berechnungen, die damit verbunden sind. Denn nur mit Hochleistungscomputern lassen sich neue Funktionen künftig in Serie bringen. „Autonomes Fahren gehört zu den anspruchsvollsten KI-Anwendungen überhaupt“, erklärt 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.“

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

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

50-mal so groß wie ein Grafikprozessor

Allerdings wurden Grafikkarten ursprünglich nicht für den KI-Einsatz entworfen, sondern um Bilddaten möglichst effizient zu verarbeiten. Sie stoßen darum beispielsweise beim Training von Algorithmen für das autonome Fahren zunehmend an ihre Grenzen. Für noch schnellere Berechnungen ist deshalb auf KI spezialisierte Hardware erforderlich. Eine mögliche Lösung hat die kalifornische Firma Cerebras präsentiert. Ihre „Wafer Scale Engine“ (WSE) ist optimal auf die Anforderungen Neuronaler Netze zugeschnitten, indem sie möglichst viel Rechenkraft auf einem riesigen Computerchip vereint. Er ist mehr als 50-mal so groß wie ein normaler Grafikprozessor und bietet Platz für 850.000 Rechenkerne – über 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 häufig vor, dass bei Vektor-Matrix-Multiplikation mit der Zahl Null multipliziert wird. Das kostet unnötig Zeit. Die Wafer Scale Engine unterlässt es deshalb. „Alle Nullen werden herausgefiltert“, schreibt Cerebras in seinem Whitepaper zur WSE. Der Chip führt 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 Wasserkühlung erforderlich. Cerebras hat darum ein eigenes Server-Gehäuse für den Einsatz in Rechenzentren entwickelt. In den Datenzentren einiger Forschungsinstitute wird die Wafer Scale Engine bereits getestet. Der Riesenchip aus Kalifornien könnte auch die Automobilentwicklung beschleunigen, glaubt Experte Schaper. „Mit ihm ließe sich theoretisch eine Woche Training auf nur wenige Stunden reduzieren“, schätzt er. „Das muss die Technik in Praxistests jedoch erst noch unter Beweis stellen.“

Licht statt Elektronen

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

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

Dafür nutzen Lightelligence und Lightmatter das Phänomen der Interferenz. Wenn sich Lichtwellen gegenseitig verstärken oder auslöschen, 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 können also „rechnen“. Um das praktisch umzusetzen, haben die Bostoner Entwickler winzige Lichtleiter in einen Silizium-Chip geätzt. Wie in einem Textilgewebe überkreuzen 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 ausführen.

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

Mit der Kombination aus optischen und digitalen Rechnern lassen sich DNNs extrem schnell berechnen. „Ihr Hauptvorteil liegt in der geringen Latenzzeit“, erklärt Lindsey Hunt, Sprecherin von Lightelligence. Das DNN kann dadurch beispielsweise schneller Objekte auf Bildern erkennen, wie etwa Fußgänger und e-Scooter-Fahrer. Beim autonomen Fahren könnte dies zu schnelleren Reaktionen in kritischen Situationen führen. „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 Lösungen von Lightmatter und Lightelligence lassen sich als Module in herkömmliche Computer einsetzen, um die KI-Berechnungen zu beschleunigen – ähnlich wie Grafikkarten. Im Prinzip könnte man sie auch in Fahrzeuge integrieren, etwa um autonome Fahrfunktionen zu realisieren. „Unsere Technologie ist sehr gut geeignet, um als Inferenzmaschine für ein autonomes Auto zu dienen“, erklärt Lindsey Hunt. Experte Schaper sieht das ähnlich: „Wenn es Lightelligence gelingt, automobiltaugliche Komponenten zu bauen, könnte dies die Einführung komplexer KI-Funktionen im Fahrzeug stark beschleunigen.“ Die Technik ist inzwischen marktreif: Das Unternehmen plant erste Pilotversuche mit Kunden für 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 können. Sie arbeiten dafür mit sogenannten „Qubits“. Anders als die klassische Informationseinheit, das Bit, kann ein Qubit

die beiden binären Werte 0 und 1 simultan darstellen. Die beiden Zahlen koexistieren in einem Überlagerungszustand, wie er nur in der Quantenmechanik möglich ist.

Quantencomputer könnten beim Klassifizieren von Dingen zum Turbo für Künstliche Intelligenz werden, zum Beispiel im Verkehrsgeschehen. Dort gibt es viele unterschiedliche Objekt-Kategorien, darunter Fahrräder, Autos, Fußgänger, Schilder, nasse und trockene Straßen. Sie unterscheiden sich anhand vieler Eigenschaften, weswegen Experten von „Mustererkennung in hochdimensionalen Räumen“ sprechen.

„Je komplizierter die Muster, desto schwerer tun sich herkömmliche Rechner damit, Klassen zu unterscheiden“, erklärt Heike Riel, die IBMs Quantenforschung in Europa und Afrika leitet. Denn mit jeder Dimension wird es aufwendiger, die Ähnlichkeit zweier Objekte auszurechnen: Wie ähnlich sind sich ein e-Scooter-Fahrer und ein Rollatornutzer, die die Straße überqueren wollen? Quantenrechner können im Vergleich zu klassischen Computern effizient in hochdimensionalen Räumen arbeiten. Für gewisse Probleme könnte diese Eigenschaft nützlich sein und dazu führen, dass mithilfe von Quantencomputern einige Probleme schneller gelöst werden können als mit klassischen Hochleistungsrechnern.

IBM-Forscher haben statistische Modelle analysiert, die sich für die Datenklassifizierung trainieren lassen. Erste Resultate deuten darauf hin, dass geschickt gewählte Quantenmodelle für gewisse Datensätze besser funktionieren als klassische Methoden. Die Quantenmodelle sind einfacher zu trainieren und scheinen eine größere Kapazität zu haben – was ihnen erlaubt, kompliziertere Zusammenhänge zu lernen.

Mit den heutigen Quantencomputern kann man diese Algorithmen testen, aber noch keinen Vorteil gegenüber 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 dürfte es gelingen, für bestimmte Anwendungen die Überlegenheit von Quantencomputern zu zeigen – also dass sie Probleme schneller, effizienter und präziser lösen als ein klassischer Computer. Einen leistungsstarken, fehlerkorrigierten, universellen Quantenrechner zu bauen, dauert aber noch etwas länger. Nach Schätzung von Experten werden dafür noch mindestens zehn Jahre vergehen. Doch das Warten könnte sich lohnen. Wie auch optische Chips oder neue Architekturen von elektronischen Rechnern könnten Quantenrechner der Schlüssel sein auf dem Weg zur Mobilität der Zukunft.

Zusammengefasst

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

Info

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

Autor: Christian Meier

Copyright: Alle in diesem Artikel veröffentlichten Bilder, Videos und Audio-Dateien unterliegen dem Copyright. Eine Reproduktion oder Wiedergabe des Ganzen oder von Teilen ist ohne die schriftliche Genehmigung der Dr. Ing. h.c. F. Porsche AG nicht gestattet. Bitte kontaktieren Sie newsroom@porsche.com für weitere Informationen.

MEDIA ENQUIRIES



Frederic Damköhler

Senior Manager Corporate Communications Porsche Engineering
+49 (0) 711 / 911 16361
frederic.damkoehler@porsche.de

Linksammlung

Link zu diesem Artikel
<https://newsroom.porsche.com/de/2022/innovation/porsche-engineering-quantencomputer-computerchip-optische-rechner-27646.html>

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