INSTITUT FÜR
INNOVATION UND
TECHNIK

iit-kompakt 03 ▪ Korbinian Schreiber

Neuromorphe Chips
Künstliche Gehirne aus Silizium?

Obwohl Computer viel schneller und genauer rechnen können als wir, haben wir ihnen doch in vielerlei Hinsicht einiges voraus. Beispielsweise lenken wir riesige SUVs durch den Feierabendverkehr, durchschauen komplexe politische Zusammenhänge oder schicken Roboter zum Mars. Verdanken tun wir diese Fähigkeiten unserem Gehirn, das unsere Gedanken mit Hilfe von ca. 100.000.000.000 Nervenzellen auf geordnete Bahnen lenkt. Mit neuromorphen Chips versuchen Forscher:innen, genau diese biologischen Funktionsweisen und Abläufe nachzuahmen um damit intelligente Computersysteme für die Zukunft zu entwickeln.

In Anlehnung an die Biologie: Chips wie Nerven

Neuromorphe Chips bestehen wie die digitalen Prozessoren in unseren Smartphones und Notebooks aus Silizium und ihre Funktion wird durch die zahlreichen mikroskopischen Schaltelemente darauf bestimmt. Der wesentliche Unterschied zu normalen Prozessoren liegt jedoch in ihrem logischen und strukturellen Aufbau, der sogenannten Chiparchitektur. Diese ist bei neuromorphen Chips neuromorph – also nervenähnlich. Gemeint ist damit, dass wie die Nervenzellen im Gehirn, zahlreiche, einfache Recheneinheiten auf dem Chip existieren, die komplex miteinander verschaltet werden können. In Anlehnung an die Biologie nennt man diese Einheiten daher Neuronen und die Verbindungen zwischen ihnen Synapsen. Im Gegensatz zu einem normalen Computerprozessor verfügt ein neuromorpher Chip nicht über eine kleine Anzahl von komplexen und großen Rechenkernen, sondern über hunderte bis hunderttausende vergleichsweise kleine, künstliche Neuronen. Diese einfachen Einheiten verarbeiten im Gegensatz zu herkömmlichen Prozessoren zudem keine abstrakten Programme, sondern verwandeln lediglich eine bestimmte Art von Eingangssignalen in eine Art von Ausgangssignalen. Diese Ausgangssignale gelangen dann über künstliche Synapsen wieder als Eingangssignale zu anderen Neuronen. Auf diese Weise entsteht ein künstliches neuronales Netzwerk (KNN).

Ein Meilenstein für die Datenverarbeitung: Künstliche neuronale Netze (KNN)

KNN haben in den vergangenen zehn Jahren disruptive Veränderungen in allen Bereichen der elektronischen Datenverarbeitung ausgelöst, da sie in der Lage sind, hochdimensionale Datensätze zu analysieren und Muster darin zu erkennen und wiederzugeben. Diese KNN sind typischerweise in Software umgesetzt und benötigen grundsätzlich keine speziellen Mikrochips um ausgeführt zu werden, sondern "nur" sehr leistungsfähige Computer mit modernen Graphikkarten (GPUs). Trotzdem treten hierbei vor allem bei großen und komplizierten Netzwerken erhebliche unerwünschte Leistungsverluste auf, die rudimentär aus den nicht an die neuronalen Strukturen angepassten Chip- und Systemarchitekturen resultieren. Um diese Verluste zu minimieren und einen möglichst natürlichen Prozessablauf zu gewährleisten, entwickeln Forscher:innen und Firmen neuromorphe Hardware.

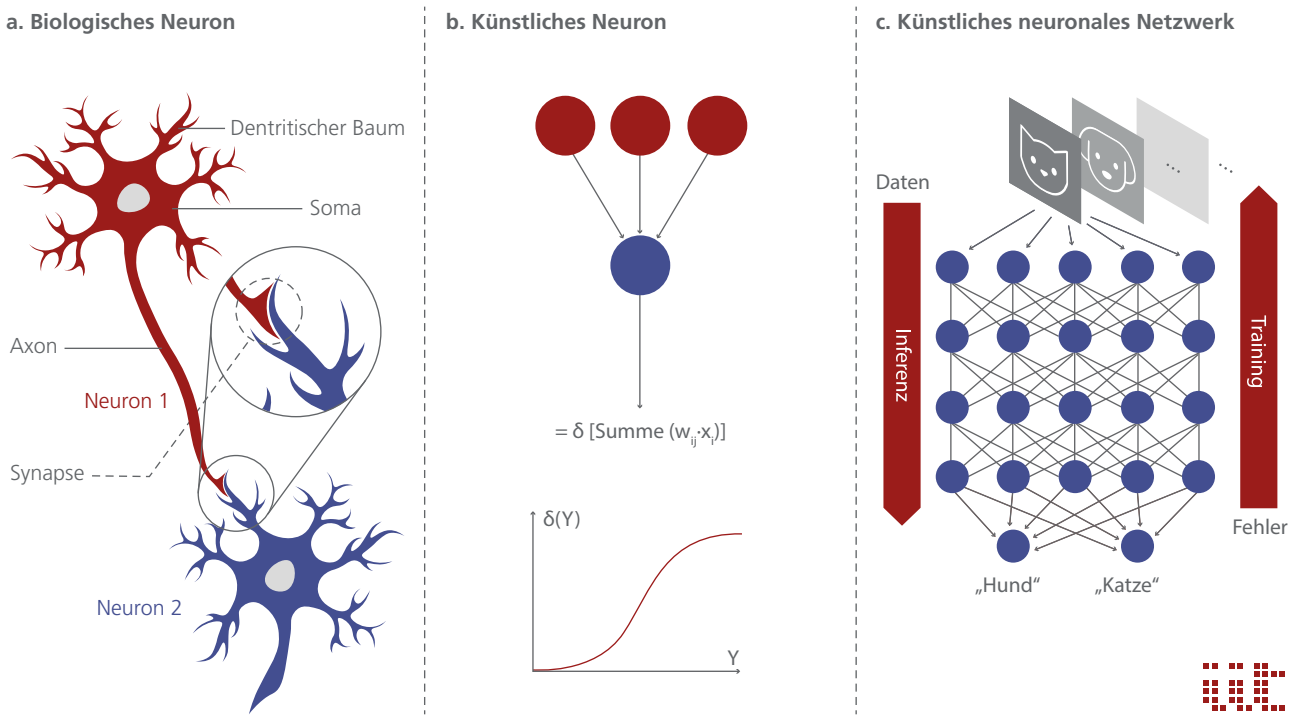

Abbildung 1: Funktionsweise und Aufbau von biologischen und künstlichen neuronalen Netzwerken. a) Zwei biologische Neuronen und eine synaptische Verbindung. b) Mathematisches Modell eines künstlichen Neurons. c) Ein künstliches neuronales Netzwerk. Eingangssignale (hier die Pixelwerte der Bilder von Hunden und Katzen) werden durch ein mehrschichtiges Netzwerk geleitet.

Prozessorarchitekturen

Von-Neumann-Prozessoren und neuroinspirierte Informationsverarbeitung

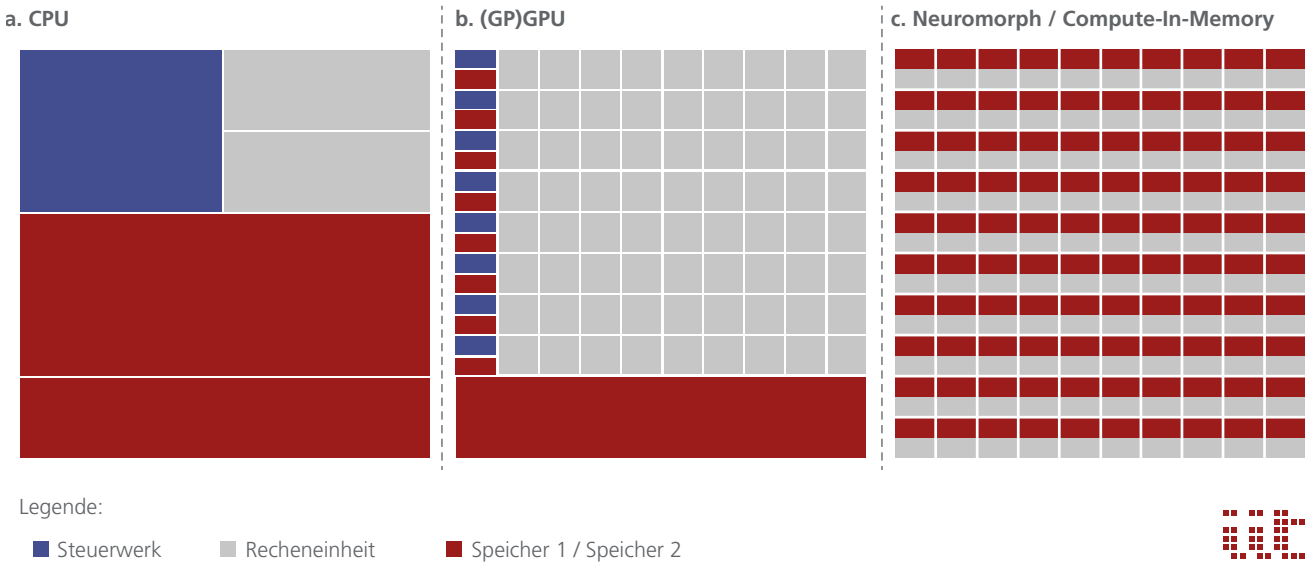

Abbildung 2: Prozessorarchitekturen. a) Architektur eines herkömmlichen Prozessors mit Steuerwerk, Recheneinheiten und Speicher. b) Architektur einer (GP)GPU mit mehreren parallel arbeitenden Steuerwerken und Speichern, sowie zahlreichen parallel arbeitenden Rechenwerken. c) Eine fiktive neuromorphe Architektur mit parallelen, mit Speichern versehenen Recheneinheiten. Sofern nur eine Art von Berechnung ausgeführt wird, kann unter Umständen sogar auf ein Steuerwerk verzichtet werden.

Von Übersetzungen bis zur Medizin: Künstliche Neuronale Netze im Alltag

Die Anwendungs- und Forschungsbereiche von KNN sind sehr vielfältig und breit gestreut, lassen sich aber grob in zwei Bereiche gliedern: das anwendungsorientierte maschinelle Lernen und die akademische Erforschung neuronaler und lernender Systeme. Die erste Art von anwendungsbezogenen KNN begegnet uns inzwischen schon täglich. KNN werden hier bereits großflächig eingesetzt um beispielsweise Konsumentenverhalten vorherzusagen, Daten (Bilder, Videos, Text) zu klassifizieren, oder um Texte zu übersetzen. Auch in anderen Bereichen, wie etwa Fahrerassistenzsystemen, der medizinischen Bildgebung und -Analyse, oder zur Spracherkennung werden KNN eingesetzt. Die dabei verwendeten Algorithmen und Architekturen entstammen dabei häufig dem akademischen Bereich. KNN wurden zunächst der beobachteten Struktur von biologischen Nervensystemen nachempfunden und über die Zeit hinweg nach immer funktionaleren Kriterien weiterentwickelt. Aktuell arbeiten Forscher:innen an einer Vereinigung der Erkenntnisse, die einerseits aus dem Anwendungsbezug des maschinellen Lernens entspringen und andererseits aus der biologischen Grundlagenforschung[1]. In den computergestützten Neurowissenschaften werden auch hierzu große Rechenleistungen benötigt und gleichzeitig Systeme entwickelt, die weitere Fortschritte in diese Richtungen ermöglichen.

Viel Leistung braucht viel Energie

Sowohl im Forschungs- als auch im Anwendungsbereich werden also große Rechenkapazitäten benötigt, die hohe Anforderungen an derzeit existierende Computer stellen. Von zentraler Bedeutung ist hierbei vor allem der Energieverbrauch. Bei großen und aufwendigen KNN limitiert dieser zuweilen die maximal erreichbare Größe des Netzwerks, da beispielsweise schlicht nicht ausreichend viel Energie zur Verfügung gestellt werden kann, oder weil der Energiebedarf bei der Kommunikation zwischen den einzelnen Rechenknoten schlecht skaliert[2]. Bei kleineren KNN, die nicht in großen Rechenzentren, sondern in kleinen Computern oder sogar in mobilen Geräten eingesetzt

1 Der Fokus liegt dabei häufig auf sogenannten Lernregeln. Damit sind formale Algorithmen gemeint, die auf lokaler, zellulärer Ebene wirken und ein Lernverhalten des Gesamtsystems verursachen. Beispielsweise zählt hierzu die Propagation von Fehlersignalen und die damit verbundenen Angleichungen der Netzwerkparameter durch das gesamte KNN (error backpropagation), oder die mathematische Modellierung von synaptischen Mechanismen, die beim Lernen in biologischen Nervensystemen auftreten.

2 Die Netzwerktopologien von KNN sind oft sehr komplex und hochdimensional, sodass eine Verdoppelung der Anzahl von Rechenknoten oft eine höhere Vervielfachung der zur Vernetzung benötigten Datenverbindungen nötig macht. Dies hat zur Folge, dass die Verbindungsressourcen für KNN nicht unbedingt linear mit der Netzwerkgröße skalieren sondern stattdessen überproportional.

werden, ist der Energieverbrauch ebenfalls für die erreichbare Rechenleistung maßgeblich und bestimmt darüber hinaus Aspekte wie Akkulaufzeit oder Hitzeentwicklung. Um an diesen Stellen für Entspannung und Fortschritt zu sorgen, werden in Form von neuromorphen Computern spezielle Mikrochips und Rechensysteme entwickelt, die mehr Rechenleistung bei gleichzeitig niedrigerer elektrischer Eingangsleistung bereitstellen.

Aktuelle Herausforderung: die Kommunikation von Chip zu Chip

Derartige, speziell angepasste Beschleunigerchips sind bereits in zahlreichen Ausführungen kommerziell erhältlich und ermöglichen für KNN-Anwendungen erhebliche Effizienzvorteile. Die verbleibenden Leistungsverluste werden dabei jedoch nicht nur durch die Berechnungen selbst bestimmt, sondern zunehmend auch durch den notwendigen Datentransfer. Der Energiebedarf für die Kommunikation eines KNN-Beschleunigers mit dem Speicher oder mit anderen Rechenknoten befindet sich inzwischen häufig auf der gleichen Größenordnung wie der für die neuromorphen Berechnungen selbst. In einigen Fällen übersteigt er diesen sogar deutlich. Aus diesem Grund verfolgen viele aktuelle Forschungsprojekte und Start-ups die großskalige Integration mehrerer Chips oder Teilsysteme zu größeren Einheiten. Neue Aufbau- und Verbundmethoden, wie dreidimensionale Integration oder on-silicon photonics, werden dazu in der nahen Zukunft einen zunehmend wichtigeren Beitrag leisten, da sie in der Lage sind, die Kommunikationsbandbreite an der Chipgrenze drastisch zu erhöhen und damit genau an der Stelle ansetzen, die aktuell den bedeutendsten Flaschenhals[3] darstellt. Optimierungen und Entwicklungen finden aber nach wie vor auch bei den Rechenwerken, also bei den neuromorphen Schaltkreisen selbst statt. Neben rein digitalen Konzepten werden auch analoge CMOS-Schaltungen, oder sogar neuartige Substrate, wie z. B. Memristoren oder photonische Medien, aktiv erforscht. In einigen Fällen werden hier vorteilhafte oder günstige Eigenschaften in Bezug auf die Energie- und Platzeffizienz vermutet und daher in neuromorphe Schaltungen integriert.

System- und Chiparchitektur als Erfolgsfaktor

Wegen einer notwendigen Anbindung solcher Chips zu konventionellen Speichermedien sind die diskutierten, konventionellen Schnittstellen jedoch auch hier unvermeidbar. Analoge Elektronik und/oder neuartige Substrate werden daher auf absehbare Zeit vor allem in kleinen Systemen für mobile oder sehr spezielle Anwendungen, bei denen der Einfluss des Datenverkehrs nicht überwiegt, nennenswerte Einsparungen erzeugen können. Für große Systeme bleiben weiterhin die Datenschnittstellen von zentraler Bedeutung.

In beiden Fällen wird die Energieeffizienz jedoch am entscheidensten durch die Architektur des neuromorphen Chips und Systems geprägt, da diese die größte Hebelwirkung auf die Steuerung, Verteilung und Taktung von Daten und Rechenprozessen ausübt. So erweisen sich Effizienzvorteile durch nichtdigitale Implementierungsarten (analog, memristiv, photonisch, etc.) gegenüber den Kommunikationsstrukturen und der unvermeidbaren digitalen CMOS-Peripherie zunehmend als zweitrangig, zumal inzwischen auch volldigitale Rechenwerke mit erstaunlichen Effizienzeigenschaften demonstriert werden konnten. Analoge oder memristive Rechenmethoden verkomplizieren aber tendenziell die Entwicklung und Benutzung neuromorpher Systeme und wirken sich damit nachteilig auf Entwicklungszeit und Marktakzeptanz aus. Digitale CMOS-Schaltungen bieten daher in dieser Hinsicht signifikante Vorteile, die wegen der fortlaufenden Miniaturisierung von Baugruppen auf absehbare Zeit Bestand haben werden[4].

Neben dem Chip selbst benötigt ein neuromorphes System auch eine umfangreiche und aufwendig gepflegte Softwareumgebung, eine benutzerfreundliche Dokumentation, und nicht zuletzt eine erfolgreiche Vermarktungsstrategie. Der finanzielle und personelle Aufwand für diese Aspekte übertrifft nicht selten den eigentlichen Chip- und Hardwareentwicklungsaufwand.

---

3 Zur Verarbeitung von Daten müssen diese grundsätzlich in den Chip hinein und in verarbeiteter Form aus dem Chip heraus gelangen. Gerade bei datenintensiven Berechnungen, wie eben bei der neuronal-inspirierten Informationsverarbeitung, stellt daher die Schnittstelle nach außen einen wichtigen Beitrag dar. Im klassischen Computing lässt sich die Schnittstelle zwischen Prozessor und dem datentragenden Medium – also dem Speicher – außerdem direkt mit dem sogenannten von-Neumann-Bottleneck assoziieren: egal wie schnell ein Prozessor ist, wenn alle Daten erst durch dieses Nadelöhr gelangen müssen, wird die gesamte Rechengeschwindigkeit letztich durch die Kapazität dieser Schnittstelle begrenzt sein. Eine wichtige Strategie im neuromorphen Computing besteht daher darin, Speicher und Prozessor zu vereinigen um so zumindest einen Teil der Daten – nämlich die synaptischen Gewichte – nicht transportieren zu müssen. Falls die Größe des zu berechnenden KNN jedoch die Rechenressourcen auf einem neuromorphen Chip übersteigt, oder falls die Menge an Eingangs- und Ausgangsdaten zur Verarbeitung sehr groß ist, kann die Bandbreite nach außen auch hier den limitierenden Faktor darstellen. Bei allen Anwendungen, die die Kapazität eines einzelnen neuromorphen Chips übersteigen, sind die Schnittstellen nach außen daher von tragender Bedeutung.

4 Obwohl sich auf der Chipebene zunehmend ein Ende des Mooreschen Gesetzes [https://de.wikipedia.org/wiki/Mooresches_Gesetz] abzeichnet, bietet die Integration einzelner Chips zu größeren Systemen noch erhebliches Potential für weiteren Effizienzgewinn. Die großen Halbleiterhersteller gehen daher für die nächsten beiden Dekaden überwiegend weiterhin von einer exponentiellen Effizienzsteigerung von CMOS-basierten Halbleitern aus.

Herausgeber

Prof. Dr. Volker Wittpahl
Institut für Innovation und Technik (iit)
in der
VDI/VDE Innovation + Technik GmbH
Steinplatz 1, 10623 Berlin

Autor

Korbinian Schreiber
Tel: 030 310078-5432
E-Mail: Korbinian.Schreiber@vdivde-it.de

iit-kompakt Nr. 03

Oktober 2022
Layout: VDI/VDE-IT

ISBN: 78-3-89-750246-8