1. Hardware für KI

Markus Schürholz, Eike-Christian Spitzner

Die KI ist bereits seit Jahrzehnten ein Thema in der Forschung, wobei die Konferenz „Dartmouth Summer Research Project on Artificial Intelligence" im Jahr 1956 als Startpunkt systematischer Forschungsanstrengungen gilt. Den wirklichen Durchbruch brachte allerdings erst in den vergangenen Jahren der Einsatz von künstlichen neuronalen Netzen (KNN) mit Methoden des tiefen Lernens (Deep Learning, DL), welche rudimentär Abläufe im Nervensystem nachbilden (siehe auch Einleitung Teil A). Wichtige Treiber sind aber nicht nur die Konzepte der KNN, sondern vor allem auch die Entwicklung der Rechentechnik, auf der entsprechende Verfahren ausgeführt werden. Während man zu Beginn auf leistungsfähige Allzweckprozessoren (central processing unit, CPU) zurückgriff, werden seit einigen Jahren vorrangig Prozessoren verwendet, die ursprünglich für Grafikkarten zur Bildausgabe gedacht waren (graphics processing unit, GPU). Aktuell werden diese zunehmend zu Spezialprozessoren (application-specific integrated circuit, ASIC) für KI-Anwendungen weiterentwickelt. Zusätzlich verfolgt man den Ansatz, die Struktur von KNN direkt in der Architektur eines Prozessors abzubilden (neuromorphe Hardware). Dabei sind erste Versuche erfolgversprechend.

Um die Entwicklung der Hardware für KI-Anwendungen besser einordnen zu können, ist es zunächst hilfreich sich anzusehen, welche Berechnungen bei der Nutzung von KNN mit DL-Ansätzen durchgeführt werden. Hierbei muss man noch klar zwischen dem Anlernen des KNN (Training) und seinem späteren Einsatz (Inference) unterscheiden, wobei ersteres sehr rechenaufwendig ist. Die in diesem Beitrag beschriebene Hardware dient insbesondere der Beschleunigung des Trainings. Im Prinzip bestehen KNN aus einzelnen konzeptionellen Neuronen, die in bestimmten Schichten angeordnet sind. Bei mehrschichtigen Netzwerken ist die erste Schicht die Eingabeschicht, die Daten entgegennimmt. Die letzte Schicht, welche das Ergebnis liefert, ist die Ausgabeschicht. Gibt es zwischen Ein- und Ausgabeschicht weitere Schichten (Hidden Neurons), wird das neuronale Netzwerk deutlich leistungsfähiger, und man spricht von DL. Zwischen den einzelnen Schichten bestehen Verbindungen zwischen Neuronen, die das eigentliche Netzwerk bilden. Diese Verbindungen haben verschiedene Strukturen, nach denen neuronale Netze auch klassifiziert werden können (siehe auch Einleitung Teil A „Entwicklungswege zur KI"). Ein einfacher Fall ist dabei ein Feedforward-Netz, in dem jedes einzelne Neuron einer Schicht über Verbin-

V. Wittpahl (Hrsg.), *Künstliche Intelligenz*,
DOI 10.1007/978-3-662-58042-4_2, © Der/die Autor(en) 2019

dungen die Informationen den Neuronen der nächsten Schicht senden, jedoch nicht zurücksenden kann.

Das eigentliche „Wissen" des Netzes steckt, entsprechend einem biologischen neuronalen Netz, in der Gewichtung der einzelnen Verbindungen zwischen den künstlichen Neuronen. Diese Struktur muss zunächst erzeugt werden, das Netz wird also angelernt. Eine gängige Methode hierfür ist das Überwachte Lernen (Supervised Machine Learning). Dabei trainiert man das Netz mit bekannten Eingangsdaten sowie Ausgangsdaten und stellt die Gewichtung der einzelnen Verbindungen so ein, dass Fehler am Ausgang minimal ausfallen. So kann ein neuronales Netz zum Beispiel trainieren, auf Bildern Hunde und Katzen zu unterscheiden, indem man am Eingang Bilder verwendet, von denen bekannt ist, welche der beiden Tierarten darauf zu sehen ist (Wert am Ausgang). Die Trainingsphase ist abgeschlossen, wenn das neuronale Netz mit unbekannten, nicht für das Training verwendeten Daten eine Fehlerrate erreicht, die unter einem vorher festgelegten und der Anwendung angemessenem Wert liegt. Grundsätzlich kann man sagen, dass ein neuronales Netz mit mehr Schichten und mehr Neuronen, zusammen mit möglichst vielen Trainingsdaten, theoretisch die besten Resultate erzeugt, gleichzeitig aber mit der Anzahl der Neuronen, der Anzahl der Schichten und der Menge an Trainingsdaten der Rechenaufwand erheblich steigt. Diese Berechnungen können auf unterschiedliche Art und Weise in Software umgesetzt werden. Wichtig dabei ist jedoch, dass die Berechnungen in der Regel so implementiert sind, dass mathematisch hauptsächlich Matrixmultiplikationen und Vektoradditionen durchgeführt werden. Im Folgenden wird am Beispiel der Matrixmultiplikation gezeigt, warum dies einen entscheidenden Einfluss darauf hat, welche Hardware für KI-Anwendungen besonders effizient ist.

Matrix A multipliziert mit Matrix B ergibt dabei eine neue Matrix C (siehe Abbildung 1.1). Die vier Elemente der Ergebnismatrix C werden dabei unabhängig aus Elementen der Matrizen A und B berechnet und enthalten keine unmittelbaren Abhängigkeiten untereinander. Das heißt, die Matrixmultiplikation kann sehr einfach in vier Rechnungen aufgeteilt werden, die nicht aufeinander aufbauen und aus diesem Grund gleichzeitig ausgeführt werden können, ohne auf ein anderes Zwischenergeb-

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} * \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{pmatrix}$$

Abbildung 1.1: Multiplikation zweier Matritzen

nis warten zu müssen. Jede einzelne Rechnung besteht dabei nur aus einer Addition zweier Multiplikationen, zum Beispiel $A_{11}B_{11} + A_{12}B_{21}$, wobei die beiden Multiplikationen auch gleichzeitig ausgeführt werden können, um in einem zweiten Schritt addiert zu werden. Die auf den ersten Blick recht aufwendige Multiplikation zweier Matrizen lässt sich so in viele einfache Teile zerlegen. Es wird deutlich, dass in einem ersten Schritt acht Multiplikationen gleichzeitig und in einem zweiten Schritt vier Additionen gleichzeitig ausgeführt werden können. Insgesamt lässt sich diese Rechnung also sehr gut parallelisieren, was wiederum der entscheidende Punkt für die Wahl der Hardware ist. Zur Verfügung stehen dafür im Allgemeinen Universalprozessoren (CPU), Beschleunigerkarten, die im Wesentlichen auf Grafikprozessoren basieren (GPU), und anwendungsspezifische Schaltungen (ASIC).

Aktuelle Hardware-Lösungen

Die meisten heute verwendeten Universalprozessoren, wie beispielsweise die Hauptprozessoren in allen gängigen Computern wie auch Mobilgeräten und Servern, basieren grundlegend auf einer Architektur, die John von Neumann im Jahr 1945 beschrieb und die auch nach ihm benannt ist (von-Neumann-Architektur). Kennzeichen dieser Architektur ist ein gemeinsamer, zentraler Speicher für Daten und Instruktionen. Dies ist konzeptionell sehr effizient, da möglichst leistungsfähige Rechenwerke die Programme sequenziell, also Schritt für Schritt, abarbeiten sollen. Optimiert ist ein solcher Prozessor für aufeinander aufbauende, komplexe Berechnungen, nicht jedoch für parallelisierbare Aufgaben. Dies gilt grundsätzlich, ist heute jedoch nur noch eingeschränkt gültig, da sich die Entwicklung der CPUs in den vergangenen Jahrzehnten ein Stück weit von den Ursprüngen entfernt hat. Moderne CPUs verfügen über hohe Taktraten und eine hohe Rechenleistung pro Takt, und durch Befehlserweiterungen sind sie in der Lage, auch komplexere Berechnungen in einem oder sehr wenigen Schritten auszuführen. Zudem ist mit diesen modernen CPUs inzwischen auch ein paralleles Abarbeiten mehrerer Aufgaben möglich, da sie mehrere Prozessorkerne (in Smartphones aktuell bis zu 10, in Serverprozessoren 32 und mehr) beinhalten und Technologien wie SMT (simultaneos multithreading) dies unterstützen – eine Technik, die es erlaubt, im begrenzten Umfang zwei Aufgaben auf demselben Prozessorkern auszuführen. Moderne CPUs sind also sehr leistungsfähig, vielseitig und können komplexe Probleme schnell bearbeiten. Für Rechnungen, die massiv parallelisiert werden können und aus eher einfachen Teilaufgaben bestehen, ist eine CPU jedoch weiterhin eher ungeeignet. Die Teilschritte werden zwar sehr schnell ausgeführt, die Anzahl der parallel ausgeführten Aufgaben ist jedoch begrenzt. Die große Rechenleistung der einzelnen Kerne und viele Optimierungen moderner Prozessoren wie etwa Befehlssatzerweiterungen können kaum oder nicht genutzt werden – mit der Folge, dass letztlich ein solcher Prozessor mit parallelen Rechenarbeiten nicht optimal ausgelastet werden kann.

In den vergangenen Jahren wurde deshalb für solche Berechnungen immer häufiger Hardware verwendet, die eigentlich für die Bildausgabe entwickelt wurde. Diese basiert auf sogenannten GPUs. Die Leistungsfähigkeit dieser Grafikhardware ist, besonders im Vergleich zu CPUs, in jüngster Zeit verhältnismäßig stark gestiegen. GPUs bestehen aus ähnlichen Einzelbausteinen wie CPUs, unterscheiden sich in der Gesamtarchitektur jedoch deutlich. Für die Berechnung einzelner Bildpunkte nutzten GPUs früher kleine Rechenkerne, sogenannte Shader, die auf bestimmte Funktionen optimiert waren und nur diese ausführen konnten. Es gab spezialisierte Shader, beispielsweise um die Farbe, die Transparenz oder Geometrie einzelner Bildpunkte oder Bildbereiche zu berechnen. Ob die einzelnen Funktionen jedoch genutzt wurden, hing dabei stark von der Software ab. Um die Hardware generell besser auslasten zu können, basieren moderne GPUs deswegen auf universellen Shadern, sogenannten Unified Shader-Architekturen. Diese generalisierten Shader sind in der Lage, je nach Bedarf jede der gewünschten Funktionen auszuführen. Bedingung ist, dass jeder Shader direkt programmiert werden kann, was ihn zu einem kleinen Universalprozessor macht. Diese Fähigkeit ermöglicht es nun, solche GPUs nicht mehr nur zur Bildberechnung zu nutzen, sondern sie auch andere Berechnungen anstellen zu lassen, was sie zu GPGPU („general purpose computation on graphics processing unit“) werden lässt. Bei der Verwendung als GPGPU kann nun jeder Shader als eine Art Universalrechenkern angesehen werden. Ein solcher Kern ist für sich genommen im Vergleich zu einem CPU-Kern zwar erheblich schwächer und deutlich niedriger getaktet, moderne GPUs verfügen jedoch über tausende entsprechender Shader, zwei Größenordnungen mehr als eine CPU. Ein weiterer Unterschied zur CPU besteht darin, dass der Speicher einer Grafikkarte um etwa einen Faktor zehn schneller angebunden ist, was besonders bei großen Datenmengen von Vorteil ist.

Eine dritte Möglichkeit Berechnungen durchzuführen, ist die Verwendung anwendungsspezifischer integrierter Schaltkreise (ASIC). Hierbei handelt es sich im Gegensatz zu CPUs und in Grenzen GPUs nicht um Universalprozessoren, die prinzipiell in der Lage sind, fast jede Berechnung durchzuführen. ASICs sind speziell für nur eine bestimmte Aufgabe entworfene Schaltkreise. Die Grenze, an der ein modifizierter oder ergänzter Universalprozessor aufhört und ein ASIC beginnt, ist dabei durchaus fließend, für die Auswahl von KI-Hardware aber nicht zwingend wichtig.

Relevant für die KI-Anwendung ist zum einen Hardware, die auf Matrixrechenoperationen spezialisiert ist. Derartige Hardware ist zurzeit in Form von speziellen, zusätzlichen Rechenkernen auf KI-Beschleunigern wie Nvidia Tensor Core (NVIDIA TESLA V100 GPU ARCHITECTURE) oder ganzen darauf spezialisierten Prozessoren wie bei Google, tensor processing unit, TPU verfügbar. Zum anderen gibt es auch Bestrebungen für KI-Anwendungen, bei denen ein KNN komplett in Hardware abgebildet werden soll, sogenannte neuromorphe Hardware.

Die aktuell gängigen Implementationen von KNN basieren darauf, dass im Wesentlichen sehr viele Matrixoperationen ausgeführt werden. Wie am Beispiel der Matrixmultiplikation gezeigt, sind solche Aufgaben inhärent parallelisierbar, lassen sich also in viele recht einfache Rechnungen zerlegen, die größtenteils gleichzeitig stattfinden können. Von den Optimierungen moderner, auch leistungsfähiger CPUs mit ihrer noch begrenzten Fähigkeit zum Parallelrechnen kann solch eine Anwendung allerdings kaum profitieren. Vielmehr können GPUs, ursprünglich für Grafikhardware bzw. Beschleunigerkarten entwickelt, hier ihr Potenzial voll ausspielen. Dies ist auch der wesentliche Grund dafür, dass viele KI-Anwendungen erst mit der Nutzung von GPUs den Durchbruch schafften. Zuvor waren nur sehr teure Großrechner in der Lage, entsprechende Berechnungen in angemessener Zeit durchzuführen. Großes Zukunftspotenzial haben auch auf Matrixoperationen spezialisierte ASICs, wie sie gegenwärtig schon nach und nach zum Einsatz kommen. Die Unterschiede in der Effizienz sind dabei deutlich: So gibt Google für die eigens entwickelte TPU – ein ASIC für Vektoroperationen – bei KI-relevanten Berechnungen etwa die 80-fache Rechenleistung gegenüber einer CPU und die 30-fache Rechenleistung gegenüber einer GPU an, wobei diese Werte auf die aufgenommene elektrische Leistung, also pro Watt, normiert sind (Jouppi et al. 2017; Hot Chips 2017: A Closer Look At Googles TPU v2).

Die skizzierten Unterschiede in den Prozessor-Architekturen verdeutlichen, welch wichtige Rolle der verwendeten Hardware für den Erfolg von KI-Konzepten zukommt. Im folgenden Abschnitt wird deshalb ein genauerer Überblick gegeben, welche Akteure hier mit welcher Hardware im Markt aktiv sind. Grundsätzlich lässt sich festhalten, dass sich die Rechentechnik für KI-Anwendungen immer weiter von der klassischen von-Neumann-Rechenmaschine entfernt. Ein interessanter Aspekt der Entwicklung, denn von Neumann hatte für sein Konzept der Rechenmaschine eigentlich das zentrale Nervensystem des Menschen durchaus als ein Vorbild betrachtet und die Gemeinsamkeiten und Unterschiede in seinem Buch „Die Rechenmaschine und das Gehirn“ (Neumann 1960) schon vor Jahrzehnten präzise durchdacht.

Marktübersicht

Zahlreiche Hersteller bieten bereits für KI-Anwendungen optimierte Rechenhardware an und es kommt stetig neue hinzu. Die erste große wirtschaftliche Erfolgsgeschichte einer KI-Hardware ist mit dem Namen Nvidia Corporation verbunden: Das in Kalifornien beheimatete Unternehmen wurde 1993 gegründet und begann mit der Kommerzialisierung von GPUs, die sich speziell für den Einsatz in der 3D-Computergrafik eigneten und mit denen sich zahlreiche Aspekte computergenerierter Bilder parallel rechnen ließen. Um die Jahrtausendwende hatte sich das Unternehmen in diesem Bereich sehr erfolgreich am Markt positioniert. Es folgten Firmenübernahmen und

Expansion, u. a. auch durch den Zukauf der Berliner Mental Images GmbH im Jahr 2007. Im gleichen Jahr veröffentlichte Nvidia mit CUDA (Compute Unified Device Architecture) eine Schnittstelle für seine Hardware, um GPGPU für das unspezifische Abarbeiten parallelisierbarer Rechenaufgaben zu ermöglichen.

Das war der Startschuss für eine breite Nutzung der Grafikkarten für DL in einer großen Forschungsgemeinschaft. Ebenfalls 2007 brachte Nvidia den ersten Prozessor der Tesla-Reihe auf den Markt, dessen aktuelle Version Volta heißt. Die Strukturgröße der Transistoren im Volta ist nur noch zwölf Nanometer groß, und der Chip umfasst mehr als 5.000 Shader – ein großer Unterschied also zu den 28 Rechenkernen in Intels aktueller CPU. Nvidia spricht in Hinblick auf die aktuellste Volta-Generation von neuen „Tensor Cores“[6]. Der Begriff in der Benennung von Chips soll darauf hindeuten, dass Matrixoperationen auf diesen Chips sehr effizient durchgeführt werden können. Während bei CPUs die Leistungszuwächse (oft beschrieben durch das „Mooresche Gesetz“) in den vergangenen Jahren von Generation zu Generation eher kleiner wurden, konnten Nvidias GPUs in den aktuellsten Generationen enorme Leistungssprünge verzeichnen.

Gegenüber CPUs, die sich seit vielen Jahren in PCs, Servern – heute meist Cloud genannt – und mittlerweile insbesondere in Smartphones befinden, konnte Nvidia mit seinen neuen KI-Chips ein völlig neues Marktsegment erschließen. Dies spiegelt sich deutlich in der unterschiedlichen Entwicklung der Aktienkurse von Nvidia und vom Hersteller klassischer CPUs Intel wider (siehe Abbildung 1.2). Und Nvidias KI-Chips können auch in der Cloud als mächtige KI-Rechencluster genutzt werden. Interessanterweise arbeitet das Unternehmen für dieses Angebot mit Microsoft und dem im Cloud-Computing dominanten Amazon zusammen. Im Rahmen seines „AI Lab“-Programms kooperiert Nvidia mit wichtigen KI-Forschungseinrichtungen. Als einen der beiden ersten europäischen Partner wählte Nvidia das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) in Saarbrücken (Auel 2016).

Aufgrund der absehbar auch künftig dynamischen Marktentwicklung von KI für eine steigende Anzahl von Anwendungen hat auch der Konzern Google, der sich die Entwicklung von KI seit Unternehmensgründung als langfristiges Ziel auf die Fahnen geschrieben hatte, eine eigene Hardware entwickelt. Deren Name TPU (Tensor Processing Unit), orientiert sich an den Begriffen CPU und GPU. Die gegenwärtig bereits in der zweiten Generation verfügbaren Google-TPUs dienen ebenfalls dazu, Matrix-

[6] *Da auch Google den Begriff Tensor für die eigene Hardware verwendet, sei kurz darauf hingewiesen, dass es sich bei einem Tensor um ein mathematisches Objekt handelt, das in einfachen Fällen eine Zahl oder ein Vektor ist, in komplexeren Fällen eine multidimensionale Matrix.*

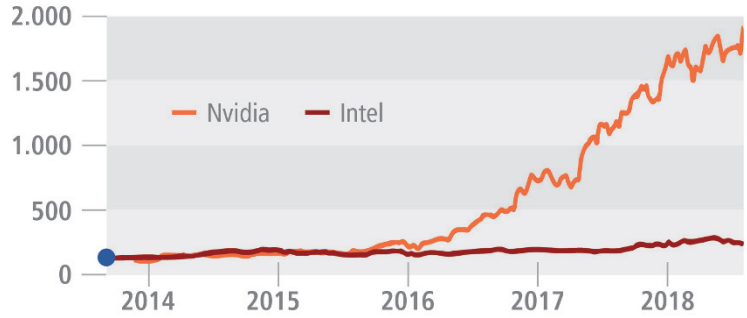

Abbildung 1.2: Aktienpreise, der Preis vom 1. Januar 2012 entspricht 100, um das Verhältnis der Kurssteigerung abzubilden (eigene Darstellung basierend auf IDC, Thomson Reuters).

operationen effizient auszuführen. Die Chips wurden dabei so gestaltet, dass die von Google entwickelte Open-Source-Softwarebibliothek TensorFlow effektiv damit verwendet werden kann. Google stellt die TPUs im Rahmen des eigenen Cloud-Angebotes zur Verfügung; prominent eingesetzt wurde die Hardware bei dem 2016 Aufsehen erregenden Sieg von AlphaGo über den Go-Spieler Lee Sedol.

Während diese Entwicklungsansätze von KI-Hardware einerseits auf den lokalen Einsatz zielen und andererseits aufgrund ihrer Effizienz mit CPUs in Rechenzentren oder Supercomputern konkurrieren, werden schon mobile Chips mit Recheneinheiten ausgestattet, die ML unterstützen. Anwendungen fallen dabei in vielen Fällen in den Bereich Computer Vision, in dem mit ML eindrucksvolle Erfolge erzielt werden konnten. Microsoft setzt beispielsweise in seiner für Augmented bzw. Mixed Reality Anwendungen entwickelten HoloLens eine Holo Processing Unit ein, die CPU und GPU unterstützt – also eine HPU, der allgemeinen Bezeichnungstradition folgend.

Gegenwärtig weitverbreitet ist der sogenannte A11 Bionic Chip, der im iPhone 8 (Plus) und X eingesetzt wird. Die System-on-a-Chips (SoCs), die bisherige iPhone-Generationen antrieben, enthielten bereits mehrere Prozessoren, neben einer CPU und GPU auch gesonderte Prozessoren, die nur Bewegung erfassen und dabei besonders energieeffizient sind. Seit dem A11 Bionic umfasst der Chip auch einen von Apple als Neural Engine bezeichneten Prozessor, der für Machine Learning insbesondere im Bereich Computer Vision angewendet wird. So ermöglicht diese Neural Engine die nahezu in Echtzeit stattfindende Entsperrung des Smartphones durch lokal ausgeführte Gesichtserkennung (Face ID). Und obwohl auch andere Hersteller von Smartphone-Chips auf lokale KI-Hardware setzen, sticht die Neural Engine auch deshalb hervor, weil sie dabei hilft, den von Apple favorisierten Entwicklungsansatz zu unterstützen, Daten so weit wie möglich auf dem Endgerät des Nutzers zu belas-

sen und dort zu verarbeiten. Während bei Google eingestellte Bilder in die Cloud geladen werden und erst dort Mustererkennung auf den Fotos stattfindet, ermöglicht die Neural Engine eine effiziente Mustererkennung von Fotos auf dem iPhone.

Ebenfalls für den Bereich Computer Vision vorgesehen ist die Vision Processing Unit (VPU) von Intel, die aktuell den Namen Myriad X trägt und auf Technologie von Movidius fußt. Bevor dieses Unternehmen 2016 von Intel übernommen wurde, stellte es die kleine und energieeffiziente Computer-Vision-Technologie für Drohnen von DJI bereit. Mit einem Verbrauch im Bereich von einem Watt eignet sich der aktuelle Myriad X für den mobilen Einsatz und kann Stereo-Bildquellen mit einer Auflösung von 720 Pixel bei einer Frequenz von 180 Hertz auswerten. Im selben Jahr wie Movidius übernahm Intel 2016 auch Nervana Systems, deren Technologie im aktuellen Nervana Neural Network Processor (NNP) verbaut wird und für den nicht-mobilen Einsatz konzipiert ist. Die beiden Übernahmen wirken wie ein Doppelschlag, um sich gegen bereits etabliertere Konkurrenten am Markt zu positionieren. Darüber hinaus übernahm Intel im Bereich Automotive das israelische Unternehmen Mobileye, das spezifische Sensoren für Fahrassistenzsysteme anbietet. Der milliardenschwere Kauf besiegelte den größten Exit der israelischen Technologiewirtschaft.

Neben Nvidia und den bekannten Riesen erforschen und entwickeln diverse Start-ups eigene Lösungen von unterschiedlicher öffentlicher Transparenz, die hier nur exemplarisch vorgestellt werden können. Zu nennen wäre beispielsweise Graphcore, ein 2016 in Großbritannien gegründetes Start-up, das sein System Intelligence Processing Unit (IPU) nennt und damit nach eigenen Angaben beeindruckende Performances erreicht. Das 2013 in Beijing gegründete Unternehmen Bitmain Technologies entwickelt ASICs, die für das Mining von Bitcoins optimiert sind. Bitmain weitet seine Aktivitäten gerade in den Bereich ASICs für KI-Anwendungen aus und verfolgt dabei technisch einen ähnlichen Ansatz wie Google. Die Lösung von Wave Computing wird Dataflow Processing Unit genannt und ist für den Einsatz in Servern bzw. der Cloud konzipiert. Wie konkurrenzfähig Start-ups wie Groq, Cerebras (beide USA) oder Cambricon (China) in der nächsten Zeit sein werden, ist noch nicht abzuschätzen.

Ausblick

Die Entwicklung von KI-Anwendungen und deren praktische wie wirtschaftliche Bedeutung werden auch künftig maßgeblich von Entwicklungen im Bereich der Hardware abhängen. Die Adaption von KNN auf GPU-Hardware war in der Vergangenheit ein essenzieller Schritt, um deren Berechnung um Größenordnungen zu beschleunigen und Zeitskalen zu erreichen, die eine praktische Anwendung erlauben. Ähnliche Schritte sind auch in Zukunft zu erwarten. Mobile KI-Anwendungen, bei denen neuronale Netze auf kleinen, mobilen Geräten ausgeführt werden, benö-

tigen Spezialhardware, die neben hoher Leistung auch eine sehr niedrige Leistungsaufnahme aufweist. Erste Entwicklungen zeigen sich etwa im Bereich der Mobiltelefone, wo KI-Koprozessoren verwendet werden, um beispielsweise die Qualität der damit aufgenommen Fotos und/oder deren inhaltliche Auswertung zu verbessern. Enorme Potenziale für die Zukunft lassen sich in aktuellen Forschungsergebnissen zu neuromorphen Prozessoren erkennen. IBM zum Beispiel zeigt bereits die zweite Generation seines Demonstrations-KI-Prozessors TrueNorth, welcher in Hardware eine Million Neuronen mit 256 Millionen Synapsen implementiert (Merolla et al. 2014). Dieser Prozessor ist in der Lage, typische Aufgaben der Bildauswertung mit hoher Genauigkeit und Geschwindigkeit durchzuführen, benötigt dafür aber im Vergleich zum kommerziellen Stand der Technik Größenordnungen weniger elektrische Energie (25 bis 275 Milliwatt) (Esser et al. 2016).

Die Hardware ist dabei deswegen so effizient, weil sie in Grenzen das KNN bereits in ihrer Schaltung widerspiegelt. Einzelne Rechenkerne bilden die Neuronen, die untereinander vernetzt sind (Synapsen), wobei jeder dieser „neurosynaptischen" Rechenkerne seinen eigenen Speicher hat. Hier zeigt sich in besonderem Maße die Abkehr von klassischen Architekturen, bei denen Rechenwerke und Speicher klar getrennt sind. Bei Berechnungen können jedoch alle Kerne mehr oder minder parallel arbeiten und blockieren sich nicht gegenseitig bei der Abfrage von Gewichtungsinformationen, die bei klassischen Architekturen in einem gemeinsamen zentralen Speicher liegen würden. Auch arbeiten die einzelnen Kerne nicht nach einem festen Takt, sondern nur, wenn sie durch relevante Aktivität anderer Rechenkerne angeregt werden, was die Effizienz erheblich verbessert und der Arbeitsweise des Gehirns ähnelt. Perfekt ist diese Technik allerdings nicht. So kann der TrueNorth-Chip ein künstliches neuronales Netzwerk nicht trainieren, sondern ist dabei auf klassische Hardware angewiesen (Honey 2018). Auch können wegen der deutlich abweichenden Hardware nicht alle Softwarewerkzeuge benutzt werden, welche sich in der Zwischenzeit etabliert haben. Nichtsdestotrotz sind erste Ergebnisse zu neuromorpher Hardware vielversprechend. Bevor es aber zu einer Verdrängung der zurzeit dominierenden KI-Hardware auf Basis von Grafikprozessoren und zum Teil ASICs kommt, müssen sicherlich noch einige Jahre Entwicklungsarbeit investiert werden. Unerwartete Effekte, wie zum Beispiel die aktuelle Knappheit und der erhebliche Preisanstieg bei Grafikprozessoren durch den Boom von Kryptowährungen wie Bitcoin und Ethereum können die Geschwindigkeit der Entwicklung jedoch durchaus beeinflussen.

Betrachtet man die aktuellen Marktteilnehmer und die sich abzeichnenden Entwicklungen im Bereich der Hardware für KI-Anwendungen, so wird deutlich, dass Know-how und Gewinne sich gegenwärtig in den USA konzentrieren und zusätzliche Akteure in China sichtbar werden. Kommerzielle deutsche Angebote finden sich gegenwärtig nicht. Dies ist eigentlich verwunderlich, denn in Deutschland sind mit der Automobilindustrie und dem Maschinen- und Anlagenbau vielversprechende KI-

Anwenderbranchen stark verankert. Branchengrößen wie Bosch und Continental setzen beispielsweise aktuell auf Chips von Nvidia. In der Grundlagenforschung zeigt sich hingegen ein anderes Bild. An der Universität Heidelberg etwa hat die Gruppe um den Physiker Karlheinz Meier den neuromorphen Hochleistungscomputer BrainScaleS entworfen und realisiert, der vier Millionen Neuronen mit einer Milliarde Synapsen in Hardware abbildet (Kerstin Sonnabend 2016; Schiermeier und Abbott 2016). Dieser Computer wird genutzt, um im Rahmen des Human Brain Projects der Europäischen Union (Human Brain Project) Vorgänge im Gehirn zu simulieren.

Literatur

Auel, Kersten (2016): Deep Learning: Nvidia kooperiert mit dem DFKI. Online verfügbar unter https://www.heise.de/ix/meldung/Deep-Learning-Nvidia-kooperiert-mit-dem-DFKI-3247792.html, zuletzt geprüft am 18.07.2018.

Honey, Christian; Waldrop, Mitchell (2018): Wettrennen um das künstliche Gehirn. Online verfügbar unter https://www.heise.de/tr/artikel/Wettrennen-um-das-kuenstliche-Gehirn-3996587.html, zuletzt geprüft am 21.03.2018.

Esser, Steven K.; Merolla, Paul A.; Arthur, John V.; Cassidy, Andrew S.; Appuswamy, Rathinakumar; Andreopoulos, Alexander et al. (2016): Convolutional networks for fast, energy-efficient neuromorphic computing. In: Proceedings of the National Academy of Sciences of the United States of America 113 (41), S. 11441–11446. DOI: 10.1073/pnas.1604850113.

Hot Chips (2017): A Closer Look At Google's TPU v2. Online verfügbar unter http://www.tomshardware.com/news/tpu-v2-google-machine-learning,35370.html, zuletzt geprüft am 21.03.2018.

Human Brain Project. Online verfügbar unter https://www.humanbrainproject.eu/en/, zuletzt geprüft am 21.03.2018.

Jouppi, Norman P.; Young, Cliff; Patil, Nishant; Patterson, David; Agrawal, Gaurav; Bajwa, Raminder et al. (2017): In-Datacenter Performance Analysis of a Tensor Processing Unit. In: CoRR abs/1704.04760.

Kerstin Sonnabend (2016): Vom Gehirn inspiriert, 24.03.2016. Online verfügbar unter http://www.pro-physik.de/details/physiknews/9108261/Vom_Gehirn_inspiriert.html, zuletzt geprüft am 21.03.2018.

Merolla, Paul A.; Arthur, John V.; Alvarez-Icaza, Rodrigo; Cassidy, Andrew S.; Sawada, Jun; Akopyan, Filipp et al. (2014): Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. In: Science (New York, N.Y.) 345 (6197), S. 668–673. DOI: 10.1126/science.1254642.

Neumann, J. von (1960): Die Rechenmaschine und das Gehirn: Oldenbourg (Scientia Nova Series). Online verfügbar unter https://books.google.de/books?id=msjK3xRMNkAC, zuletzt geprüft am 21.03.2018.

NVIDIA TESLA V100 GPU ARCHITECTURE. Online verfügbar unter http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf, zuletzt geprüft am 21.03.2018.

Schiermeier, Quirin; Abbott, Alison (2016): Flagship brain project releases neuro-computing tools. In: Nature 532 (7597), S. 18. DOI: 10.1038/nature.2016.19672, zuletzt geprüft am 21.03.2018.