Rechnen mit Licht: Photonische Prozessoren für künstliche neuronale Netze

Von Dr. Frank Brückerhoff-Plückelmann

Der Einfluss künstlicher neuronaler Netze auf unsere Gesellschaft, Arbeitsweise und Wissenschaft wächst stetig. Konkrete Beispiele sind Sprachmodelle wie ChatGPT des US-Konzerns OpenAI, die erfolgreich zur Generierung von Text und Bildern genutzt werden und sogar komplexe Aufgaben, die logische Schlussfolgerungen erfordern, bewältigen können. Gleichzeitig eröffnen tiefe neuronale Netze wie AlphaFold von Google DeepMind, das die komplexe Struktur von Proteinen vorhersagen kann, völlig neue Möglichkeiten in Wissenschaft und Forschung. Sowohl die grundlegende Arbeit an neuronalen Netzen als auch ihr direkter Einsatz in der Wissenschaft wurden 2024 mit den Nobelpreisen in Physik und Chemie gewürdigt. Während wirtschaftliche und ethische Fragen rund um die Integration von künstlicher Intelligenz (KI) in den Alltag häufig im Mittelpunkt der öffentlichen Diskussion stehen, stellt auch die nachhaltige Bereitstellung der erforderlichen Rechenressourcen eine zentrale Herausforderung dar [1]. Ein eindrucksvolles Beispiel ist das Training eines GPT-3-Modells mit 175 Milliarden Parametern: Es benötigte auf 10.000 Nvidia V100-Grafikkarten fast 15 Tage und verbrauchte dabei 1.287 MWh [2] – das entspricht der gesamten Leistungsabgabe eines großen Kernkraftwerks über eine Stunde hinweg. Zudem besteht der Trend zu immer größeren Modellen, da sich die Anzahl der trainierbaren Parameter, vor allem im Bereich der Sprachmodelle, meist positiv auf die Performance auswirkt. Diese enormen Ressourcenanforderungen verdeutlichen, dass effizientere und leistungsstärkere Hardwarebeschleuniger benötigt werden, um die weitläufige Nutzung von KI nachhaltig zu gestalten und zudem eine weitere Entwicklung der Modelle zu ermöglichen.

Um dedizierte Hardware für künftige KI-Anwendungen zu entwickeln, lohnt sich ein Blick in die Vergangenheit. Die ersten Arbeiten zum Perzeptron, einer grundlegenden Struktur in KI-Modellen, stammen bereits aus den 1940er Jahren. Trotz anfänglicher Begeisterung und vereinzelter Durchbrüche sorgte die mangelnde Rechenleistung bald für Ernüchterung. Dies führte zu mehreren KI-Wintern, in denen Forschung und Investitionen in künstliche neuronale Netze stark zurückgefahren wurden. Das grundlegende Problem ist leicht zu verstehen: Neuronale Netze, inspiriert von ihrem biologischen Vorbild, basieren auf wenigen Operationstypen, hauptsächlich Additionen und Multiplikationen, die oft ausgeführt werden müssen. Um ein Modell effizient zu berechnen, müssen diese Operationen daher massiv parallel ausgeführt werden. Dies steht im starken Kontrast zur damals vorherrschenden Von-Neumann-Architektur, die für allgemeine Berechnungen konzipiert ist und auf einer sequenziellen Programmausführung basiert. Einen entscheidenden Durchbruch brachte Mitte der 2000er-Jahre die Idee, Grafikkarten – ursprünglich für parallele Berechnungen in Videospielen entwickelt – für künstliche neuronale Netze zu nutzen. Heute sind spezialisierte KI-Beschleuniger wie Googles Tensor Processing Units gezielt auf diese Anforderungen optimiert. Neben paralleler Berechnung liegt ihr Fokus insbesondere auf der Minimierung von Speicherzugriffen, um die Effizienz zu maximieren. Interessanterweise weisen viele dieser Hardware-Optimierungen Ähnlichkeiten zur Funktionsweise des menschlichen Gehirns auf. So dienen z. B. Synapsen in biologischen neuronalen Netzen sowohl als „Speicher“ für die Verbindungstärke zwischen zwei Neuronen als auch als „Rechner“, der die Anregung eines Neurons mit dem „gespeicherten“ Wert „verrechnet“. Im Gegensatz dazu weist die klassische Von-Neumann-Architektur eine strikte Trennung zwischen Rechen- und Speichereinheit auf. Diese Entwicklung motiviert die Erforschung neuer KI-Hardware, die noch stärker vom menschlichen Gehirn inspiriert ist.

In meiner Doktorarbeit habe ich dieses Ziel mithilfe lichtbasierter analoger Rechner verfolgt. Die beiden fundamentalen Unterschiede zu konventioneller Hardware – analoges statt digitales Rechnen und optische statt elektrische Signalverarbeitung – bieten neue Möglichkeiten für die Entwicklung neuromorpher Systeme. Analoge Rechner kodieren eine Zahl direkt durch eine physikalische Größe, etwa durch die Helligkeit eines Lichtpulses, anstatt ein abstraktes binäres Kodierungsschema zu verwenden. Dadurch lassen sich Rechenoperationen direkt auf physikalischer Ebene im Speicher durchführen und folglich energieintensiver Datentransfer vermeiden. Ein Beispiel hierfür ist ein Objekt mit veränderbarer Transparenz, das gleichzeitig als Speicher dient, indem es einen Wert in seiner Transparenz speichert, und als Recheneinheit fungiert. Wenn ein Lichtpuls durch das Objekt tritt, wird er abgeschwächt – die Ausgangshelligkeit entspricht dem Produkt aus Eingangshelligkeit und Transparenz. Die Verwendung von Licht für Berechnungen bietet besondere Vorteile, die unter anderem auch in Glasfasernetzen genutzt werden. Die hohe Bandbreite im Terahertz-Bereich ermöglicht die parallele Verarbeitung großer Datenmengen mit hoher Geschwindigkeit. Gleichzeitig sind die Propagationsverluste innerhalb des Schaltkreises erheblich geringer als in elektrischen Systemen, was den Kühlaufwand drastisch reduziert. Darüber hinaus bietet Licht zahlreiche physikalische Freiheitsgrade, die sich gezielt für neue Rechenansätze nutzen lassen. Besonders interessant ist die Verwendung optischen Rauschens, um die Unsicherheit von KI-Modellen zu erfassen. In meiner Forschungsarbeit habe ich daher eine konventionelle und eine probabilistische Rechenarchitektur entwickelt.

Einblick in KI: Was muss eigentlich berechnet werden?

Um spezielle Prozessoren für KI-Anwendungen zu entwickeln, muss man zunächst verstehen, wie neuronale Netze funktionieren. Das ist bereits für sich ein spannendes Forschungsgebiet, denn auch das menschliche Gehirn ist noch nicht vollständig erforscht – ein einfaches Nachbauen ist daher nicht möglich. In der Biologie bestehen neuronale Netze aus Neuronen, die über Synapsen miteinander verbunden sind. Wenn ein Neuron aktiviert wird, sendet es elektrochemische Impulse, sogenannte Spikes, an andere Neuronen. Die Synapsen verändern dabei die Stärke des Signals je nachdem, wie stark zwei Neuronen miteinander verbunden sind. Kommt an einem Neuron genügend Signalstärke in einem Zeitraum an, kann es wiederum aktiv werden und eigene Impulse aussenden. Viele der modernen tiefen künstlichen neuronalen Netze funktionieren ähnlich, allerdings ohne diese zeitliche Dynamik. Sie basieren auf dem *Universal Approximation Theorem*, das besagt, dass ein neuronales Netz jede mathematische Funktion nachbilden kann, wenn es ausreichend Neuronen und/oder Lagen besitzt.

Abbildung 1a zeigt den Aufbau eines einlagigen neuronalen Netzes, das aus einer Eingangslage mit vier Neuronen und einer Ausgangslage mit einem Neuron besteht. Beispielsweise könnten die Neuronen in der Eingangslage eines solchen Netzes die Helligkeitswerte der einzelnen Pixel eines Bildes aufnehmen. Mathematisch kann man sich die Werte der Neuronen in einer bestimmten Lage als Vektor mit N Komponenten vorstellen. Die Berechnung der nächsten Schicht erfolgt durch eine Matrix-Vektor-Multiplikation, gefolgt von einer nichtlinearen Aktivierungsfunktion. Das bedeutet: Hat die erste Schicht N Neuronen und die zweite ebenfalls N Neuronen, müssen N^2 Multiplikationen und N nichtlineare Operationen durchgeführt werden. Daher sind Matrix-Vektor-Multiplikationen die mit Abstand rechenintensivsten und energieaufwendigsten Operationen in neuronalen Netzen. Und daher liegt der größte Gewinn in der Entwicklung von spezialisierter Hardware, die genau diese Berechnungen effizienter und schneller ausführen kann.

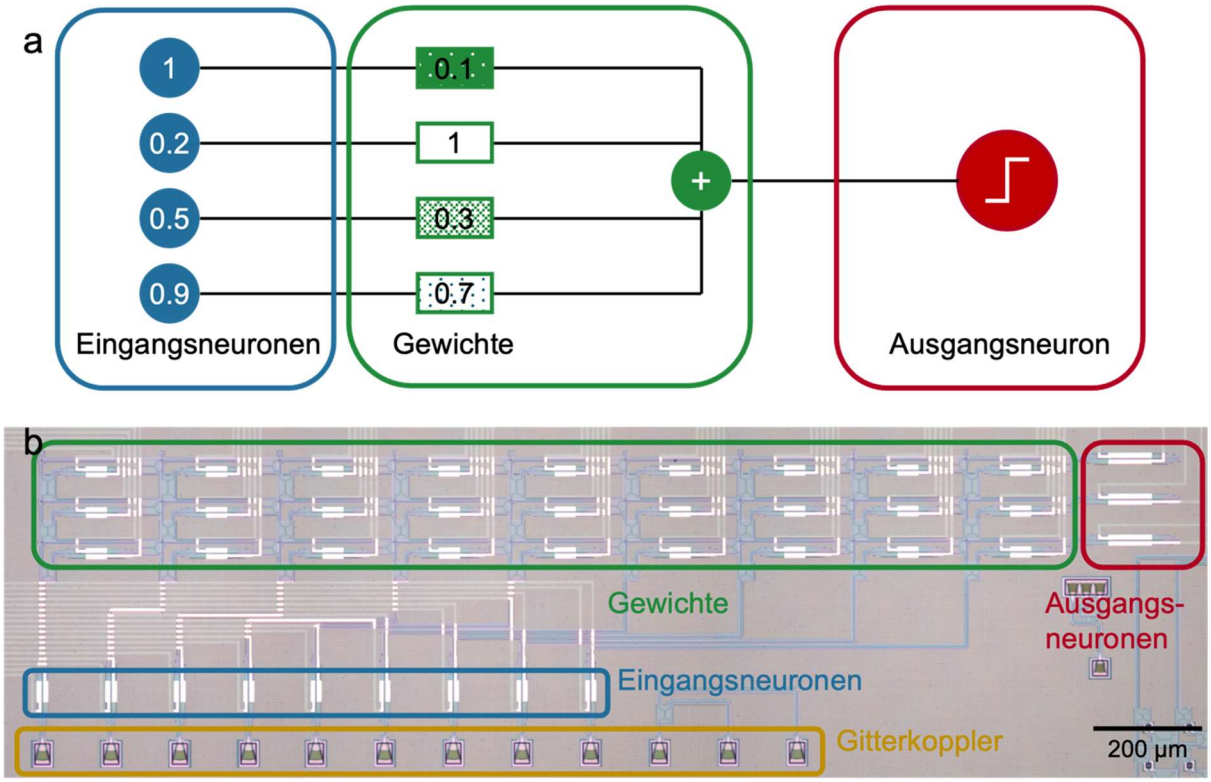

Abbildung 1. KI und photonische Prozessoren. ***a****, Um den Wert eines Ausgangsneurons zu bestimmen, wird die Aktivierung der Eingangsneuronen mit den entsprechenden synaptischen Gewichten multipliziert. Die Produkte werden addiert und die Aktivierung des Ausgangneurons durch eine nichtlineare Funktion berechnet.* ***b****, Der photonische Schaltkreis berechnet die Multiplikationen und Additionen für neun Eingangsneuronen und drei Ausgangsneuronen gleichzeitig. Die Gitterkoppler (gelb) koppeln das Licht in den Chip, die Eingangsmodulatoren (blau) kodieren die Aktivierung der Eingangsneuronen und die Photodioden messen den Wert der gewichteten Summe zur Berechnung der Ausgangsaktivierung (rot). Der Rechenspeicher speichert gleichzeitig die Gewichte und berechnet die Multiplikationen und Additionen (grün).*

In meiner Forschungsarbeit habe ich einen photonischen Schaltkreis, siehe Abbildung 1b, entwickelt, der Matrix-Vektor-Multiplikationen mit Licht berechnet [3], anstatt mit elektrischen Pulsen wie in herkömmlichen Prozessoren. Ähnlich wie in biologischen neuronalen Netzen werden dabei Informationen durch physikalische Größen dargestellt – in diesem Fall durch die Helligkeit von Lichtpulsen und die Transmission einstellbarer Absorber. Zudem werden die Rechenoperationen direkt im Speicher ausgeführt, was einen grundlegenden Vorteil gegenüber klassischen Systemen bietet. Der von mir entwickelte photonische Schaltkreis, der eine Matrix-Vektor-Multiplikation mit neun Eingangsneuronen und drei Ausgangsneuronen berechnet, besteht aus mehreren Komponenten. Die Gitterkoppler lenken Licht von einer externen Lichtquelle in den Chip ein – vergleichbar mit einem Stromkabel, das elektrische Energie in einen Computer leitet. Danach übernehmen Elektro-Absorptions-Modulatoren die Steuerung der Lichtintensität. Diese arbeiten mit Geschwindigkeiten von bis zu 50 GHz und kodieren die Werte der Eingangsneuronen direkt in die Helligkeit der Lichtpulse. Dabei entspricht eine maximale Helligkeit einem Neuronenwert von 1, während ein Wert von 0 bedeutet, dass das Licht maximal blockiert wird. Der eigentliche Rechenprozess erfolgt in einer Verschaltung verschiedener photonischer Bauteile, in Abbildung 1b in Grün umrandet. Hier werden die Addition und Multiplikation der Eingangswerte durchgeführt. Der Ablauf funktioniert wie folgt: Jeder der Eingangslichtpulse wird zunächst in drei gleichhelle Teilpulse aufgespalten. Anschließend wird die Helligkeit jedes dieser Teilpulse individuell von einem Elektro-Absorptions-Modulator

abgeschwächt. Die Stärke dieser Abschwächung entspricht den synaptischen Gewichten des neuronalen Netzes. Abschließend werden in jeder der drei Zeilen die individuell abgeschwächten Lichtpulse wieder zu einem einzigen Lichtsignal kombiniert. Dieses Signal wird schließlich mithilfe einer Photodiode ausgelesen, wodurch das Ergebnis der Berechnung ermittelt wird. Diese photonische Rechenarchitektur unterscheidet sich grundlegend von herkömmlichen elektronischen Systemen. Während konventionelle Rechner solche Multiplikations- und Additionsoperationen (sequenziell) mit Taktraten im Gigahertz-Bereich ausführen, ist die Geschwindigkeit in meinem Schaltkreis nur durch die Lichtgeschwindigkeit in den optischen Wellenleitern begrenzt.

Demonstrator: Maschinelles Sehen mit Lichtgeschwindigkeit

Der photonische Schaltkreis bildet das Herzstück des optischen Prozessors. Allerdings benötigt dieser eine Schnittstelle zur überwiegend digitalen und elektronischen Umgebung, um Daten ein- und auszulesen. Für den elektronischen Datenaustausch wird der Chip auf eine elektrische Leiterplatine geklebt. Die Kontaktpads des Chips werden über dünne Drähte mit der Platine verbunden, sodass elektrische Signale übertragen werden können (siehe Abbildung 2a). Diese Leiterplatine verbindet den Chip mit Digital-zu-Analog- und Analog-zu-Digital-Wandlern. Dadurch können digitale Daten in analoge Signale umgewandelt, im optischen Schaltkreis verarbeitet und anschließend wieder in digitale Daten zurücktransformiert werden. Die optische Kopplung erfolgt über eine Reihe von Glasfasern, die in einem Glasblock verankert sind (in Abbildung 2a rot leuchtend).

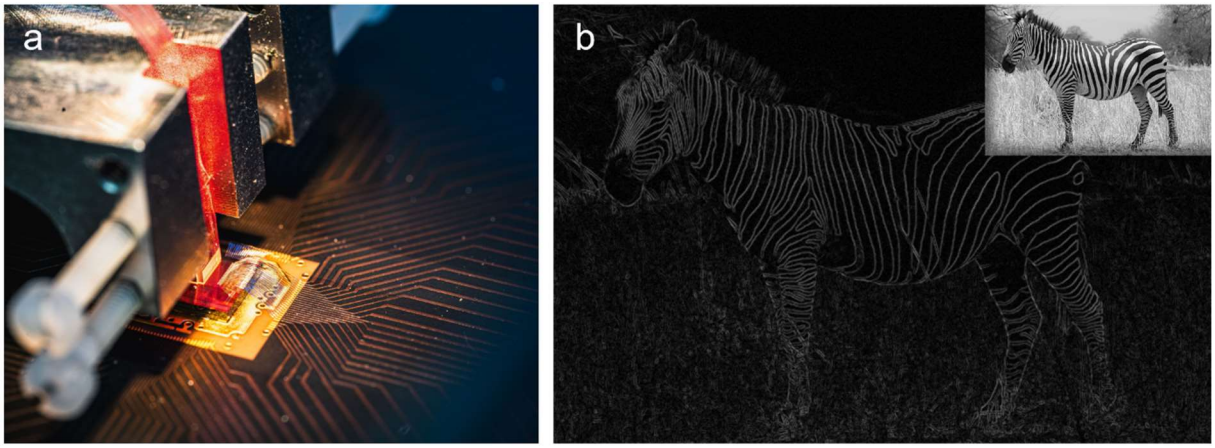

Abbildung 2. Kantenerkennung mit einem photonischen Schaltkreis. ***a****, Zur Ansteuerung des photonischen Prozessors ist der Chip auf eine Leiterplatine geklebt. So werden gleichzeitig elektrische Signale an den Chip geschickt und Licht in den Schaltkreis gekoppelt.* ***b****, Der Chip kann unter anderem Faltungen zur Kantenerkennung berechnen. Das Inset zeigt das Eingangsbild.*

Der optische Prozessor kann 2 Milliarden Matrix-Vektor-Multiplikationen pro Sekunde ausführen [3]. Seine maximale Leistung wird aktuell durch das elektronische Interface begrenzt, das für die Datenübertragung zwischen der digitalen Umgebung und dem photonischen Schaltkreis sorgt. Unter anderem im Rahmen des EU-Forschungsprojekts Phoenics habe ich diesen Demonstrator (Versuchsaufbau) genutzt, um sogenannte Faltungen – Matrix-Multiplikationen, die zur Verarbeitung visueller Daten wichtig sind – optisch zu berechnen. Bei der Faltung wird ein kleiner numerischer Filter, auch Kernel genannt, über das Bild bewegt, um lokale Merkmale wie Kanten zu identifizieren. Diese Informationen werden dann im neuronalen Netzwerk weiterverarbeitet, um beispielsweise Objekte im Bild zu klassifizieren. In Abbildung 2b ist ein Kantenerkennungsfilter auf das Bild

eines Zebras angewendet worden – alle dazu notwendigen Matrix-Vektor-Multiplikationen wurden dabei vollständig mit dem photonischen Schaltkreis berechnet. Das System haben wir sogar in einem Echtzeitszenario getestet: Ein Live-Video einer Webcam wurde direkt in den optischen Prozessor gestreamt, um die Kanten im Bild in Echtzeit zu erkennen. Dies demonstriert das enorme Potenzial optischer Berechnungen für zukünftige KI-Anwendungen, insbesondere in Bereichen wie automatisierte Bildverarbeitung und maschinelles Sehen.

Chaos nutzen: Wie optisches Rauschen KI zuverlässiger und sicherer macht

Neben der Anwendung innerhalb konventioneller neuronaler Netze eröffnen photonische Prozessoren auch neue Rechenoperationen. Diese ermöglichen es, andere mathematische Modelle effizient zu berechnen – ähnlich wie der Einsatz von Grafikkarten die Entwicklung künstlicher neuronaler Netze revolutioniert hat. Ein besonders interessanter Bereich ist die präzisere Modellierung von Unsicherheiten. Dadurch könnte eine KI nicht nur Vorhersagen treffen und Handlungsempfehlungen geben, sondern auch eine Einschätzung dazu liefern, wie sicher sie sich dabei ist. Ein typisches Beispiel ist das Verhalten in unbekannten Situationen. Menschen wissen intuitiv, wenn sie mit einer völlig neuen, nie zuvor gesehenen Situation konfrontiert sind – und dass sie in solchen Fällen keine verlässliche Aussage treffen können. Klassische KI-Modelle hingegen können leicht in die Irre geführt werden und selbst bei völlig unbekannten Eingaben mit hoher Zuversicht falsche Aussagen treffen. Gerade in sicherheitskritischen Anwendungen wie dem autonomen Fahren kann dies schwerwiegende Konsequenzen haben. Ein möglicher Lösungsansatz sind Bayessche neuronale Netze (BNNs). Anders als in klassischen neuronalen Netzen, in denen die Gewichte als feste Werte gespeichert sind, werden sie in BNNs als Wahrscheinlichkeitsverteilungen dargestellt. Dadurch kann ein solches Netzwerk nicht nur eine Entscheidung treffen, sondern auch abschätzen, wie zuverlässig diese ist. Es erkennt beispielsweise, ob es sich unsicher fühlt, weil es eine bestimmte Eingabe noch nie zuvor gesehen hat oder weil die vorhandenen Daten keine eindeutige Vorhersage zulassen. In der Praxis erfordern BNNs jedoch probabilistisches Rechnen, insbesondere die Fähigkeit, zufällige Werte gemäß Wahrscheinlichkeitsverteilungen zu ziehen. Da klassische Prozessoren deterministisch arbeiten, müssen sie mithilfe von Pseudozufallszahlengeneratoren große Rechenkapazitäten aufwenden, um die benötigten Zufallswerte zu erzeugen – ein rechenintensiver und ineffizienter Prozess. In meiner Doktorarbeit habe ich einen photonischen probabilistischen Rechner entwickelt, der diese Zufallswerte direkt während der Matrix-Vektor-Multiplikationen erzeugt. Dabei arbeitet er mit den gleichen hohen optischen Datenraten wie der zuvor vorgestellte photonische Chip für klassische neuronale Netze. Der Schlüssel dazu liegt in der spontanen verstärkten Emission innerhalb von Erbium-dotierten Glasfasern, die von Natur aus zufällige Leistungsfluktuationen erzeugen. In Kombination mit einem speziell entwickelten Kodierungsschema und einem photonischen Schaltkreis (ähnlich dem in Abbildung 1b) konnte ich damit probabilistische Matrix-Vektor-Multiplikationen berechnen. Zum praktischen Test haben wir ein photonisches BNN darauf trainiert, handgeschriebene Ziffern von 0 bis 8 korrekt zu klassifizieren. Zusätzlich wurde es mit der zuvor nie gesehenen Zahl „9“ konfrontiert – und erkannte sie erfolgreich als unbekannten Input, anstatt eine falsche Zuordnung mit hoher Zuversicht vorzunehmen [4]. Dies zeigt das Potenzial photonischer Rechner für die nächste Generation sicherer und vertrauenswürdiger KI-Systeme.

Ausblick: Photonische Prozessoren in jedem Computer?

Photonische Prozessoren bieten ein enormes Potenzial, um analoge Berechnungen effizient durchzuführen. Besonders im Bereich der künstlichen Intelligenz eröffnen sie neue Möglichkeiten, da sich ihre Architektur stärker an der Funktionsweise biologischer neuronaler Netze orientieren kann. Dadurch lässt sich die Rechenleistung steigern und der Energieverbrauch reduzieren – zwei der größten Herausforderungen aktueller KI-Hardware. In meiner Doktorarbeit habe ich Prototypen photonischer Prozessoren entwickelt, die sowohl für konventionelle neuronale Netze als auch für probabilistische neuronale Netze eingesetzt werden können. Trotz dieser Fortschritte gibt es noch Herausforderungen auf dem Weg zur breiten Anwendung photonischer Rechner. Die größte Hürde sind die hohen Entwicklungskosten, die für die Marktreife einer neuen Technologie erforderlich sind. Elektronische Architekturen profitieren von jahrzehntelangen Investitionen, während photonische Systeme erst am Anfang dieser Entwicklung stehen. Deshalb werden photonische Prozessoren zunächst vor allem in spezialisierten Hochleistungsanwendungen zum Einsatz kommen, wo der Rechenaufwand besonders hoch ist und klassische Architekturen an ihre Grenzen stoßen. Ein vielversprechendes Beispiel ist die Integration photonischer Verbindungen in Rechenzentren. Google nutzt bereits optische Verbindungen zwischen Rechenchips, um den Datentransfer in großem Maßstab effizienter zu gestalten [5]. Die Integration photonischer Komponenten auf Chip-Ebene wird intensiv erforscht, insbesondere mit Blick auf die steigenden Datenmengen, die in KI-Anwendungen verarbeitet werden müssen. Photonische Rechner könnten von diesen Fortschritten profitieren, insbesondere wenn immer mehr Peripheriekomponenten von elektrischen auf optische Schnittstellen umgestellt werden. Dadurch würde sich das Interface zwischen klassischer und photonischer Hardware vereinfachen, was eine breitere Anwendung erleichtert. Ein besonders spannendes Zukunftsfeld ist das probabilistische Rechnen, insbesondere mit Bayesschen neuronalen Netzen. Diese bieten theoretisch große Vorteile, sind aber bislang nicht weit verbreitet, da es keine wirklich effiziente digitale Umsetzung gibt. Physikalische analoge Rechner wie photonische Systeme sind hier besonders interessant, da sie natürliches Rauschen als Zufallsquelle nutzen können, anstatt sie künstlich zu erzeugen – ein entscheidender Vorteil gegenüber klassischen, rein digitalen Lösungen.

Insgesamt zeigt meine Forschungsarbeit, dass photonische Rechner ein enormes Potenzial haben, insbesondere in Kombination mit optischen Interconnects und den damit verbundenen Fortschritten in skalierbaren Fertigungsmethoden. Die Entwicklung in diesem Bereich steckt noch in den Anfängen, aber die Weichen sind gestellt – und in den kommenden Jahren könnten photonische Prozessoren eine zentrale Rolle in der nächsten Generation von KI-Hardware spielen.

Referenzen

1. Crawford, K. Generative AI’s environmental costs are soaring — and mostly secret. **Nature** 626 (2024).
2. Patterson, D., Gonzalez, J., Le, Q. et al. Carbon Emissions and Large Neural Network Training. **ArXiv** (2021).
3. Dong, B.*, Brückerhoff-Plückelmann, F.*, Meyer, L. et al. Partial coherence enhances parallelized photonic computing. **Nature** 632 (2024).
4. Brückerhoff-Plückelmann, F., Borras, H., Klein B. et al. Probabilistic photonic computing with chaotic light. **Nat Commun** 15 (2024).

5. Jouppi, N. P., Kurian, G., Li, S. et al. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. **Proc Int Symp Comput Archit** (2023).