

Physik Nobelpreis 2024 – Grundlagen der künstlichen Intelligenz

Es schlug wie eine Bombe ein, als Microsoft die von der kalifornischen Firma OpenAI lizenzierte Technik in seine Produkte integrierte. Die Öffentlichkeit war erstaunt durch die menschliche Sprache, die ChatGPT generierte. Wer die neue Technologie nicht verabscheute, versuchte, sie zur Unterstützung in sein Leben einzubauen. Der vorliegende Artikel beleuchtet die physikalischen Konzepte, die diesen Durchbruch ermöglichten, mit besonderem Augenmerk auf die Leistungen der beiden Träger des Physik Nobelpreises 2024, John J. Hopfield und Geoffrey Hinton.

Um selbst einen Eindruck vom Wesen des *Open-AI GPT-4o mini* zu erhalten, habe ich es nach einem Thema der Psychologie von C. G. Jung gefragt: «Hast du auch einen persönlichen Schatten?» In wenigen Sekunden rieselte die folgende Antwort über den Bildschirm meines Computers: *«GPT-4o mini: Als KI habe ich kein Bewusstsein, keine Emotionen und keine persönliche Identität, daher habe ich auch keinen persönlichen Schatten im psychologischen Sinne. Der Begriff «persönlicher Schatten» bezieht sich auf unbewusste Aspekte der menschlichen Psyche, die mit individuellen Erfahrungen, Emotionen und Persönlichkeitsmerkmalen verbunden sind. Da ich keine menschlichen Erfahrungen oder Emotionen habe, kann ich keine solchen Aspekte besitzen. Mein Ziel ist es, Informationen bereitzustellen und Fragen zu beantworten, basierend auf dem Wissen, das mir zur Verfügung steht.»*

Verdutzt über diese treffende Antwort verspürte ich ein grosses Verlangen, die Basis für diese technische Wunderleistung besser zu verstehen. Dabei wird sofort klar, dass Künstliche Intelligenz (KI oder engl. AI für Artificial Intelligence) auf Strukturen basieren muss, die wenig mit der Architektur herkömmlicher Computer zu tun haben. Abb. 1 zeigt die Kreativität eines anderen auf Bilder spezialisierten KI-Systems. Während herkömmliche Computermodelle wenige bis vielleicht hunderte

von Parametern besitzen, die sorgfältig kalibriert werden müssen, enthält GPT-3 (GPT steht für *Generative Pre-trained Transformer*) 175 Milliarden Gewichtsparameter, die während der Lernphase automatisch immer präziser eingestellt werden. Jeder Parameter ist 16 Bit oder 2 Byte lang (d.h. es gibt 65536 verschiedene Einstellungen) und der Speicherplatz allein für diese Parameter beträgt 350 Gigabyte.

Da ein KI-System eine gigantische Menge Stoff lernen muss und dabei dauernd der 350 Gigabyte-Speicher angepasst werden muss, ist klar, dass sehr schnelle Maschinen gebraucht werden und möglichst viele Prozesse parallel ablaufen müssen. Weiter ist verständlich, dass die gespeicherte Information (das Wissen der Maschine) über den gesamten Speicher verteilt ist und nicht wie bei normalen Computern in einzelnen Bytes steckt. So wird man Begriffe wie Relativitätstheorie oder Einstein nirgends finden, da sie nicht lokal, sondern als gegenseitige Grösse aller 175 Milliarden Gewichte gespeichert sind.

Normalerweise stürzt ein Computer ab, wenn ein einziges Memorybyte nicht korrekt funktioniert. Nicht so bei KI-Systemen: Bei einem Ausfall von Tausenden von Gewichtsbytes kann das System in seiner Genauigkeit wohl beeinträchtigt werden, aber es wird weiter funktionieren und nicht abstürzen.

Alles begann mit dem Ising-Modell zur Erklärung des Ferromagnetismus

In den 1920er-Jahren entwickelte der deutsche Physikprofessor Wilhelm Lenz (1888 - 1957) an der Universität Hamburg ein mathematisches Modell, um die Eigenschaften des Ferromagnetismus besser zu verstehen. Man kann einen Eisenstab magnetisieren, indem man ihn mit einem Draht umwickelt und darin einen starken Gleichstrom fliessen lässt, der ein Magnetfeld in Längsrichtung des Stabs erzeugt. Nach Abschalten des Stroms bleibt die Magnetisierung erhalten, falls der Eisenstab deutlich kühler ist als 761°C (Curie-Temperatur, benannt nach Pierre Curie 1859-1906). Oberhalb dieser kritischen Temperatur geht jedoch die Magnetisierung verloren. Durch Arbeiten des französischen Physi-



Abb. 1: Resultat eines auf Bilder spezialisierten KI-Systems, das am 17.8.2023 durch folgende Aufgabenstellung aktiviert wurde: «A photograph of an astronaut riding a horse». (Bild: VulcanSphere - Generated with NightCafe Studio in Stable Diffusion XL mode (1.0), wikimedia commons, Public Domain)

kers Pierre-Ernest Weiss (1865 - 1940) als Professor an der ETH Zürich von 1902 bis 1919 waren die mikroskopischen Weisschen Bezirke und die quantenhaften Eigenschaften der magnetischen Momente in den Atomen (sog. Magnetone) bekannt.

Lenz betrachtete ein zweidimensionales Gitter mit magnetischen Dipolmomenten (sog. Spins), die zur Vereinfachung nur 2 Zustände einnehmen konnten, nämlich +1 (Spin up) und -1 (Spin down). Das Faszinierende war, dass z.B. ein $N=25 \times 25$ -Gitter 2^{625} oder umgerechnet 10^{188} verschiedene Zustände einnehmen kann. Dies sind sehr viel mehr Zustände als es Atome im Weltraum gibt! Lenz fragte sich, was geschieht, wenn sich benachbarte Spins beeinflussen und formulierte eine einfache Energiefunktion, die für jeden Zustand die dazugehörige innere Energie des Systems liefert. Sie ist minimal, wenn alle Spins in dieselbe Richtung zeigen.

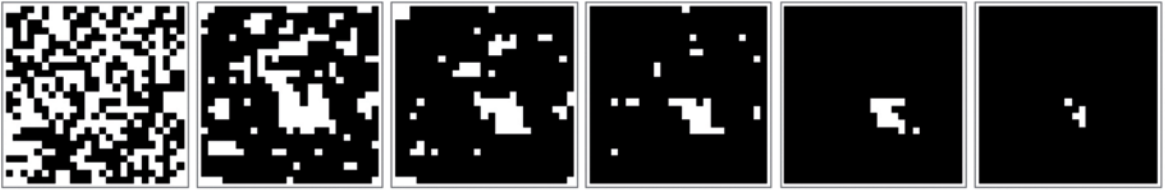
Lenz hat die Aufgabe zur Untersuchung eines eindimensionalen Modells seinem Studenten Ernst Ising gegeben und deshalb das Modell als Ising-Modell bezeichnet. Verschiedenste Physiker haben sich über mehrere Jahrzehnte mit Ising-Modellen beschäftigt und es wurde zum fundamentalen Modell der statistischen Thermodynamik. So

wurde z.B. gefunden, dass zwei- und mehrdimensionale Modelle einen Phasenübergang zeigen, wie in Abb. 2 dargestellt: Für hohe Temperaturen herrscht eine ungeordnete Phase vor (Paramagnetismus) und sobald die Temperatur eine kritische Schwelle unterschreitet, zeigt sich eine ferromagnetische Ordnung, bei der alle Spins in dieselbe Richtung zeigen.

Hopfield verwendet das Ising Modell für ein neuronales Netzwerk

Obwohl der Physiker John J. Hopfield (geb. 1933 in Chicago und tätig an der Princeton University im Bundesstaat New Jersey, USA) den sog. *Backpropagation-Algorithmus* bekannt machte, der in frühen KI-Modellen extensiv angewendet wurde, sah er in den 1980er-Jahren, dass damit nicht weiterzukommen war. In seinen Publikationen von 1982 betonte er, dass dieser zutiefst mathematische Algorithmus kaum als Modell für evolutive Nervenzentren dienen kann. Er versuchte deshalb einen Neuanfang und knüpfte an das erfolgreiche physikalische Ising-Modell an, das wesentlich einfacher aufgebaut war und trotzdem mit seinen vielen quasistabilen Zuständen kollektive emergente Phänomene zeigen konnte (vgl. Abb. 2b).

a) Unterhalb Curie-Temperatur mit Magnetfeld: **Ferromagnetismus**



b) Unterhalb Curie-Temperatur ohne Magnetfeld: **Weissche Bezirke**



c) Oberhalb Curie-Temperatur ohne Magnetfeld: **Paramagnetismus**



Abb. 2: Simulationsrechnungen mit einem zweidimensionalen Ising-Modell in einem 25x25-Gitter.

a) Ferromagnetismus: Deutlich unterhalb der Curie-Temperatur ordnet ein äusseres Magnetfeld eine zufällige Ausrichtung der Spins. Da die Energie dieser Anordnung minimal ist, bleibt die Magnetisierung auch nach Abschalten des äusseren Magnetfeldes bestehen.

b) Weissche Bezirke: Ohne äusseres Magnetfeld entstehen zusammenhängende Bezirke. Diese Klumpen-Struktur ist weitgehend stabil, da sich die Energie in einem lokalen Minimum befindet. Es gibt verschiedene solche Strukturen, die bei einem KI-Memory als Erinnerungs-Speicher verwendet werden.

c) Paramagnetismus: Wird ein magnetisierter Ferromagnet über die Curie-Temperatur aufgeheizt, zerfällt die geordnete Struktur auf Grund der thermischen Bewegungen und die Magnetisierung verschwindet.

(Bilder und Rechnungen mit Mathematica durch Fritz Gassmann)

In einer seiner 1982er-Publikationen (Hopfield 1982) findet man als Formel (7) die Kopie der Energiefunktion des Ising-Modells, obschon sein neuronales Netzwerk (es wird heute als Hopfield-Netzwerk bezeichnet) keine energetischen Zustände besass. Das Ising-Modell musste also umgedeutet und auch erweitert werden. Hopfield hat einen Muster-Speicherungs Algorithmus erfunden, der es erlaubte, vorgegebene Muster auf verblüffend einfache Weise im Netzwerk zu speichern und wieder abzurufen. Wichtig war ihm dabei, dass es so aufgebaut war, dass ein evolutives Analogon möglich erschien. Die Speicherarchitektur unterschied sich radikal von herkömmlichen Computern, wo Informationen an genau spezifizierten Adressen untergebracht werden und sich gegenseitig absolut nicht beeinflussen können.

Diese lokale Strategie hat einen grossen Nachteil: Der Ausfall einzelner Bits führt dazu, dass In-

formationen nicht mehr korrekt abgerufen werden können, oder nicht mehr gefunden werden oder dass sogar das gesamte System «abstürzt». In einem Hopfield-Netzwerk werden die Muster nicht lokal gespeichert, sondern über das ganze Netzwerk verteilt, indem der Speicher-Algorithmus die Kopplungsstärken w_{ik} zwischen den Neuronen i und k modifiziert. Ein solches Netzwerk ist relativ stabil gegenüber dem Ausfall von einzelnen Neuronen und «stürzt nicht ab».

Während normale Computer synchron und seriell funktionieren und als Taktgeber eine Uhr haben, die das gesamte System steuert (heute im Gigahertz-Rhythmus), funktioniert ein Hopfield-Netzwerk asynchron und parallel, d.h. jedes Neuron arbeitet für sich. Zugeschnitten auf die langsamen Computer der 1970er-Jahre musste sich Hopfield mit wenigen simulierten Neuronen zufrieden geben: er arbeitete mit $N = 30$ und $N = 100$ Neuronen

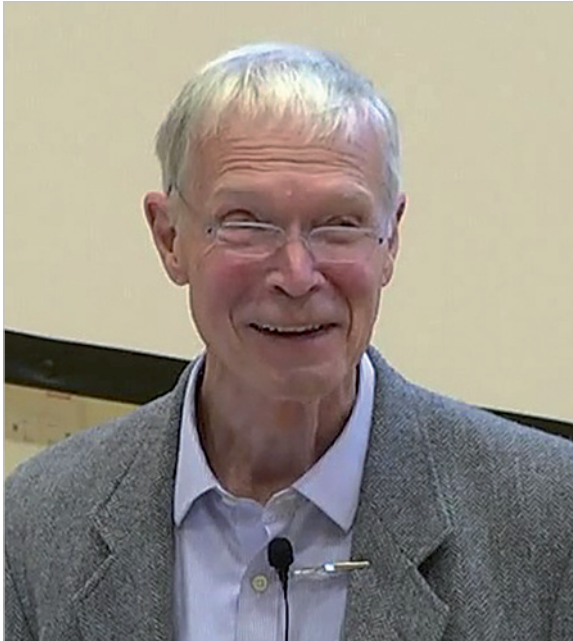


Abb. 3: Der Physiker John J. Hopfield (geb. 1933) aufgenommen 2016. Er ist sehr beunruhigt durch die neuesten Entwicklungen der KI und verglich sie mit der Entdeckung der Kernspaltung, die Nuklearenergie aber auch Atombomben ermöglichte. (Bild: bhadeshia123, wikimedia commons CC BY 3.0)



Abb. 4: Der Informatiker und Kognitionspsychologe Geoffrey E. Hinton (geb. 1947) bei einem Vortrag in Toronto 2024. Er kündigte 2023 seine Zusammenarbeit mit Google, um offen über Risiken der KI sprechen zu können. (Bild: Vaughn Ridley, wikimedia commons CC BY 2.0)

entsprechend rund 900 und 10 000 Gewichtungsfaktoren w_{ik} . Trotz dieser mikrigen Netze fand er wichtige Resultate. So konnte er rund 0,15 N verschiedene Muster speichern, also rund 5 und 16.

Mit dem 30er-Netzwerk testete er, was oberhalb der Kapazitätsgrenze passiert. Er fand, dass bei 15 (statt nur 5) aufgeprägten Mustern viele Fehler passieren, das «Gedächtnis» funktionierte jedoch immer noch und blockierte nicht mit einer Error-Meldung. Er sah auch, dass nicht nur symmetrische Gewichtsmatrizen $w_{ik} = w_{ki}$ funktionieren und konnte ohne grosse Einbussen die halbe Matrix auf Null setzen. Dieses Resultat war wichtig im Hinblick auf biologische Netzwerke, die nicht symmetrisch verbunden sind (biologische Neuronen machen nicht Synapsen vom Neuron i zum Neuron k und gleichzeitig vom Neuron k zum Neuron i).

Hinton verbessert das Hopfield Netzwerk

Der Informatiker und Kognitionspsychologe Geoffrey G. Hinton, Professor am Computer Science Department der Universität Toronto (Kanada) seit 1987, entwickelte 1985 zusammen mit Terrence

J. Sejnowski das Hopfield-Netzwerk weiter zur sog. Boltzmann-Maschine. Er behob damit einen Schwachpunkt des Hopfield-Netzwerks, das vielfach in suboptimalen lokalen Minima stecken blieb.

Wie zuvor Hopfield, liess er sich dafür beim Ising-Modell inspirieren. Dieses Modell enthielt nicht nur eine Energiefunktion, sondern verwendete auch die für viele physikalische Problemstellungen äusserst nützliche Boltzmann-Statistik. Diese wurde begründet durch den österreichischen Physiker Ludwig E. Boltzmann (1844 - 1906) und beschreibt die Energieverteilung bei Systemen mit sehr vielen Freiheitsgraden, die im Kontakt mit einem Wärmereservoir der Temperatur T stehen.

Die Boltzmann-Statistik drückt mathematisch aus, wie wahrscheinlich es ist, dass das System einen Zustand mit um dE höherer Energie einnimmt, also von der Energie E zufällig zur Energie $E+dE$ wechselt. Die Wahrscheinlichkeit für einen Zustand mit Energiezuwachs dE ist nach Boltzmann proportional zu $e^{-dE/kT}$ (k ist die Boltzmann-Konstante und T die absolute Temperatur). Damit kann man beispielsweise berechnen, wie der Luftdruck mit zunehmender Höhe abnimmt oder wieviele

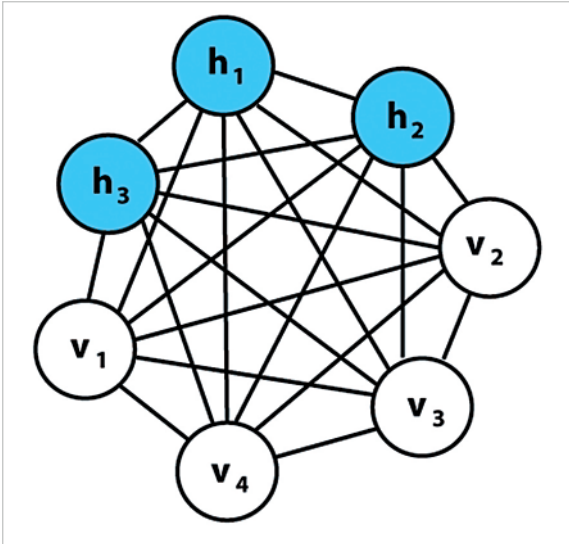


Abb. 5: Grafische Darstellung einer Boltzmann-Maschine mit 3 versteckten Neuronen (Blau) und 4 sichtbaren Neuronen (Weiss). Für Inputs (Lernen) und Outputs (Abfragen) werden nur sichtbare Neuronen verwendet. Man beachte, dass jedes Neuron mit jedem anderen gekoppelt ist. Die 7 Neuronen besitzen also $7 \times 6/2$ unidirektionale Verbindungen (Anzahl Linien), die durch je ein Gewicht w_{ik} charakterisiert werden. (Bild wikimedia commons, gemeinfrei)

Elektronen in einem Halbleiter (Diode, Transistor) das Leitungsband erreichen können.

Obwohl ein Neuronales Netz weder eine Energie E noch eine Temperatur T hat, übernahm Hinton den physikalischen Zusammenhang und bezeichnet sein Netzwerk als «Boltzmann-Maschine» (Hinton 2014). Um zu unterstreichen, dass es sich hierbei um eine Analogie zum entsprechenden physikalischen Modell handelt, spricht er nicht von Energie sondern von Kosten und statt von Temperatur spricht er lieber von zufälligen Störungen (die Boltzmann-Konstante k lässt er ganz fallen, da diese in Neuronalen Netzen keinen Sinn macht).

Hinton musste auch das Lernprozedere seiner Boltzmann-Maschine anpassen, das die Aufgabe lösen muss, die riesige Anzahl Gewichte w_{ik} (bei GPT-3 sind dies 175 Milliarden) bei jedem Lernschritt so anzupassen, dass die Kostenfunktion für den jeweils neuen Datenvektor (Muster) minimal wird. Je kleiner die Kostenfunktion, desto grösser ist die Wahrscheinlichkeit, diesen Datenvektor bei einer Suchanfrage wieder zu finden, was bedeutet, eine korrekte Antwort zu produzieren.

Da ein Neuronales Netz sehr viele Informationen (Muster) lernen muss, ist es entscheidend,

dass dieser mathematisch aufwändige Lernschritt möglichst effizient gestaltet wird. Hinton erfand dafür einen effizienten Algorithmus, den er *Contrastive Divergence* nannte (Hinton 2002). Neben weiteren Optimierungen hat Hinton die Wahrscheinlichkeiten durch deren Logarithmus ersetzt, was den Rechenaufwand deutlich reduziert (p proportional e^{-dE} wird zu $\ln(p)$ proportional $-dE$).

Heutige KI-Systeme mit vielen zusätzlichen Komponenten

Obwohl das neuronale Netz der Kern eines KI-Systems ist, sind viele zusätzliche Komponenten nötig, um das System brauchbar zu machen. So wird eine gestellte Frage durch eine Interpretationskomponente in kleinere Einheiten (sog. Tokens) zerlegt. Diese müssen dann in Muster transformiert werden, die das neuronale Netz als Input verstehen kann. Durch geeignete Steuerung wird erreicht, dass das Netz die Kostenfunktion fortlaufend weiter reduzieren kann und so von allgemeineren Aussagen (Output-Muster) zu immer spezielleren gelangt.

Diese Muster müssen anschliessend wieder in Begriffe der geforderten Sprache rückübersetzt werden. Eine Semantik- und Grammatik-Komponente formuliert in der Folge für Menschen verständliche Sätze. Falls auch ein Sprachmodul angeschlossen wird, kann man mit dem KI-System auch telefonieren. Jede dieser Komponenten ist für sich je ein komplexes System, in dem eine lange Entwicklungsgeschichte und sehr viel Know-how steckt. Die verblüffendsten Leistungen wie Flexibilität und eine Art Kreativität stammen jedoch zum grössten Teil aus dem oben erläuterten neuronalen Netz, also aus einem rein mathematisch-physikalisch beschreibbaren System.

Fritz Gassmann

Literatur

Hinton G. E. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation* 14: 1771 - 1800.

Hinton G. E. 2014. Boltzmann Machines. *Encyclopedia of Machine Learning and Data Mining*. Springer Science+Business Media New York.

Hopfield J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, Vol. 79: 2554 - 2558.