

ChatGPT & Co.

Was es mit dem neuen KI-Hype auf sich hat

Ausarbeitung zum Vortrag im Rotary-Club Frankfurt am Main Friedensbrücke

Dr. Daniel Korn

21.09.2023

1 Einleitung

Kaum ein IT-Produkt ist seit Beginn des Jahres 2023 so in aller Munde wie „ChatGPT“. Ihm wird schon jetzt die Fähigkeit zugesprochen, Menschen in vielen Bereichen, in denen es auf die Abfassung von Texten ankommt, auf absehbare Zeit zu verdrängen bzw. zu ersetzen. So wurde im Frühjahr 2023 etwa in der Presse berichtet, dass GPT 4 das US-amerikanische Staatsexamen in Rechtswissenschaften („bar exam“) mit einem sensationellen Ergebnis von mehr als 297 Punkten bestanden hat [13]. Innerhalb der letzten Monate hat sich das kostenlos nutzbare ChatGPT sehr schnell – gerade bei jüngeren Menschen – als Werkzeug zur Textproduktion in allen relevanten Lebensbereichen etabliert. Viele Nutzer lassen schon jetzt ihre täglichen E-Mails, Hausaufgaben, Studienarbeiten, Aufsätze, Artikel oder sonstige im Alltag anfallenden Schriftstücke routinemäßig von ChatGPT abfassen.

Für den unbedarften Durchschnittsmenschen erheben sich daher im Zusammenhang mit dem aktuellen Hype um ChatGPT und andere sogenannte generative KI-Systeme eine Reihe von Fragen, wie etwa die folgenden:

- Was genau ist ChatGPT eigentlich und wie funktioniert es?
- Was kann ChatGPT wirklich und wo sind seine Grenzen?
- Was haben wir von Systemen dieser Art in Zukunft zu erwarten?

Der hiesige Vortrag hat sich daher zur Aufgabe gemacht, diesen Fragen nachzugehen und ihre Beantwortung möglichst so zu präsentieren, dass sie auch für den IT-Laien einigermaßen verständlich sind. Dazu wollen wir zunächst einen kurzen Ausflug in die Welt der sogenannten künstlichen neuronalen Netze machen, um die Grundzüge der Funktionsweise von ChatGPT verstehen zu können. Darauf aufbauend wollen wir ergründen, welche Stärken sich aus dieser technischen Ausrichtung ergeben und wo sie ggf. auf ihre Grenzen stößt. Schließlich wollen wir noch einen Ausblick darauf liefern, was man in Zukunft von generativer KI nach Art von ChatGPT zu erwarten hat.

2 Künstliche neuronale Netze

In diesem Abschnitt wollen wir zunächst motivieren, welche Überlegungen zur Erfindung sogenannter künstlicher neuronaler Netze geführt haben. Anschließend wollen wir einen Blick auf die prinzipielle Funktionsweise solcher Netze werfen, bevor wir auf deren geschichtliche

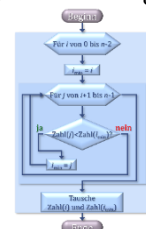
Entwicklung und die Besonderheiten eingehen, die bei der Verwendung neuronaler Netze für generative Systeme nach Art von ChatGPT zu beachten sind.

2.1 Selbstlernende vs. algorithmische Systeme

Herkömmlicher Weise werden IT-Systeme so gestaltet, dass sich der bzw. die Programmierer intensiv mit der Problemstellung auseinandersetzen, die das zu schaffende System lösen können soll. Anschließend wird ein Programm erstellt, das aus einer bestimmten Abfolge von Schritten zur Lösung des Ausgangsproblems besteht. Voraussetzung dafür ist jedoch, dass der Programmierer in der Lage ist, das Problem und seine Lösung hinreichend genau zu beschreiben.

4 3 5 1 2
1 2 3 4 5

Klassische Lehrbeispiele für diese Art des Vorgehens sind etwa Sortier- oder Suchverfahren, bei denen es darum geht, eine vorgegebene Menge von Elementen anhand eines bestimmten Kriteriums in eine geordnete Reihenfolge zu bringen bzw. ein Element mit bestimmten Eigenschaften in einer strukturierten Anordnung von Elementen zu finden. Dies sind typische Problemstellungen, die sich relativ leicht formell beschreiben lassen, so dass es grundsätzlich möglich ist, die für die Problemlösung nötigen Schritte als Folge von Anweisungen und damit als Computerprogramm zu formulieren.



3 6 8 1 7
6 7 5 7 8
2 1 7 9 7
4 8 1 9 0

Gänzlich anders geartet sind Problemstellungen, bei denen wir Menschen prinzipbedingt große Schwierigkeiten haben, hinreichend genau zu beschreiben, wie eine geeignete Problemlösung aussehen könnte. Klassische Beispiele sind all jene Aufgaben, für die auch wir Menschen üblicherweise nicht explizit lernen, wie sie zu lösen sind, sondern bei denen wir eine aus Erfahrung geprägte Intuition aufbauen. Dazu gehört insbesondere die visuelle oder akustische Erkennung bestimmter Muster (z.B. Schrift, Bildelemente, gesprochene Sprache) aber auch ebenso die kreative Erstellung von Bild-, Schrift- oder Tonwerken. So können wir Katzen von Hunden bzw. Äpfel von Birnen nicht etwa dadurch unterscheiden, dass man uns als Kinder genau erklärt hat, worin der jeweilige Unterschied zwischen ihnen besteht. Vielmehr haben wir oft genug Katzen und Hunde bzw. Äpfel und Birnen gesehen, um ein intuitives Verständnis dafür aufzubauen, wodurch sich diese Tiere bzw. Obstsorten voneinander unterscheiden lassen.



Ähnliches gilt für visuelle Texterkennung oder akustische Spracherkennung. Keiner hat uns je explizit erklärt, wodurch man einen – noch so nachlässig dahingekritzelten – Buchstaben „A“ vom Buchstaben „B“ unterscheiden kann oder welche Frequenzgänge genau den Klang des Wortes „Buchstabe“ ausmachen. Wir haben einfach oft genug Schriftzeichen in allen möglichen Variationen gesehen bzw. Sprache gehört, um ein intuitives Verständnis für die Merkmale aufzubauen, anhand derer wir Buchstaben voneinander unterscheiden bzw. Schallfrequenzgänge als Wörter einer bestimmten Sprache identifizieren können.

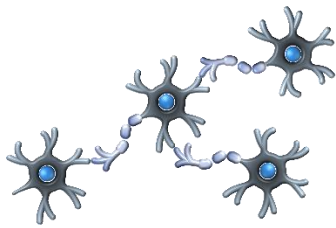
Entsprechend schlecht können wir einer Maschine explizit erklären, wie genau sie Buchstaben voneinander unterscheiden oder Schallfrequenzgänge bestimmten Wörtern einer Sprache zuzuordnen soll. Wünschenswert wäre daher, wenn wir die Maschine so konstruieren könnten, dass sie ebenso wie wir anhand verschiedenster Beispiele nach und nach selbständig lernt,

worin sich Buchstaben voneinander unterscheiden bzw. wie man einzelne Wörter einer Sprache aus Schallfrequenzgängen herausliest.

Genau hier setzen die künstlichen neuronalen Netze an, deren Funktionsweise wir im folgenden Unterabschnitt kurz umreißen wollen.

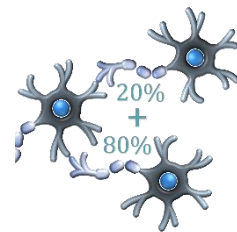
2.2 Grundzüge künstlicher Neuroner Netze

Vor dem Hintergrund der im vorangegangenen Unterabschnitt erläuterten Idee, Maschinen zu bauen, die bestimmte Inhalte wie wir Menschen selbständig anhand von Beispielen erlernen können, bietet es sich an, zunächst zu untersuchen, wie das menschliche Gehirn Informationen



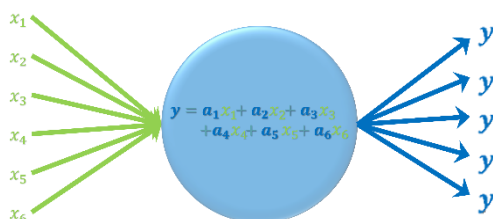
verarbeitet und speichert. Nach bisherigen Erkenntnissen wird im menschlichen Gehirn dazu eine hochgradig miteinander vernetzte Menge von rund 100 Milliarden sogenannter Neuronen verwendet. Dabei werden die Verbindungen zwischen den Neuronen durch sogenannte Axone gebildet an deren Ende elektrische Signale mit Hilfe sogenannter Synapsen auf die Membran des nächsten Neurons übertragen werden. In der Regel haben

Neurone viele eingehende Axon-Verbindungen („Dendriten“) und nur eine oder sehr wenige ausgehende Verbindungen („Neuriten“) zu anderen Neuronen. Die Frage, ob ein Neuron ein Signal an die nachgeschalteten Neuronen abgibt („feuert“), hängt dabei entscheidend davon ab, welche Signale von den Vorgängerneuronen übertragen werden. Je nach Art der Verbindung zu den Vorgängerneuronen können sich deren Signale gegenseitig hemmen bzw. pro Synapse teils stärker, teils schwächer gewichtet werden, so dass immer eine bestimmte Kombination von Eingangssignalen notwendig ist, um ein Neuron zum „feuern“ zu veranlassen.



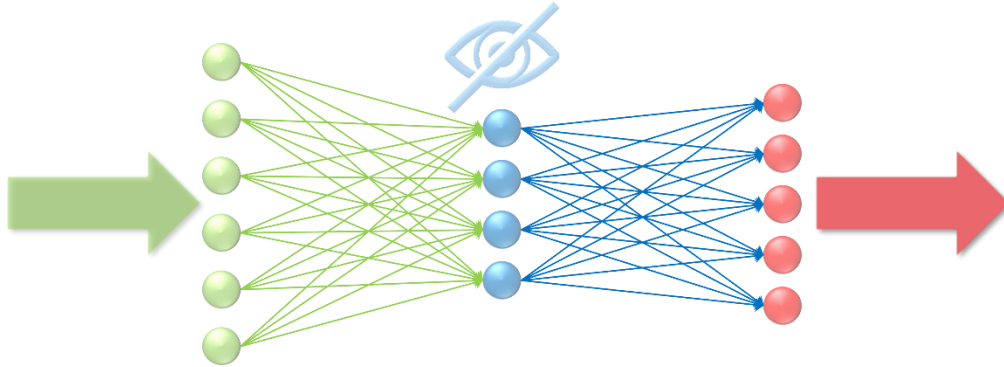
Die Frage, ob an einer Synapse eine hemmende oder verstärkende Wirkung entsteht bzw. mit welcher Gewichtung das dort übertragene Signal zum Gesamtpotenzial des Zielneurons beiträgt, wird dabei entscheidend von der Häufigkeit geprägt, mit der eine bestimmte Verbindung genutzt wird. Auf diese Weise verändert sich die Schaltlogik der Neuronenvernetzung also je nach Nutzung bestimmter Verbindungen, so dass sich die Schaltlogik also im Laufe der Zeit an den Fluss der durch sie geleiteten Signale anpassen kann. Diesem Mechanismus wird im Allgemeinen die Schlüsselrolle für die Lernfähigkeit natürlicher neuronaler Netze zugeschrieben.

Bereits in den 1940er-Jahren wurden Überlegungen angestellt, wie man die eben beschriebene Funktionsweise natürlicher neuronaler Netze auf künstliche Systeme übertragen kann. Die dazu postulierte Lösung, die sich vor allem in den 1950er Jahren konkretisierte, hat dabei die nachfolgend beschriebene grundsätzliche Gestalt:



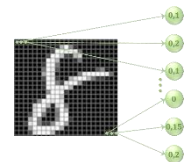
Grundelemente eines künstlichen neuronalen Netzes sind ebenfalls „Neuronen“, bei denen es sich um logische Datenstrukturen handelt, die aus einem Wert (= einer Zahl) sowie einer Reihe von Parametern zur Gewichtung ihrer Eingangssignale bestehen. Mehrere solcher Neuronen bilden dann eine sogenannte Schicht. Ein künstliches neuronales Netz besteht

schließlich aus mehreren solcher Schichten, deren einzelne Neuronen jeweils mit allen Neuronen ihrer Vorgängerschicht (sofern vorhanden) und ihrer Nachfolgeschicht (sofern vorhanden) vernetzt sind. Die erste Schicht (also diejenige ohne Vorgänger) wird als „Eingabeschicht“, die Letzte (also diejenige ohne Nachfolger) als „Ausgabeschicht“ bezeichnet. Die dazwischenliegenden Schichten bezeichnet man als „versteckte Schichten“:

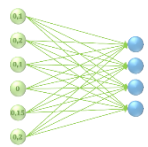


Ein so strukturiertes Netz wird dann wie folgt zur Berechnung von Ausgabewerten anhand vorgegebener Eingabewerte verwendet:

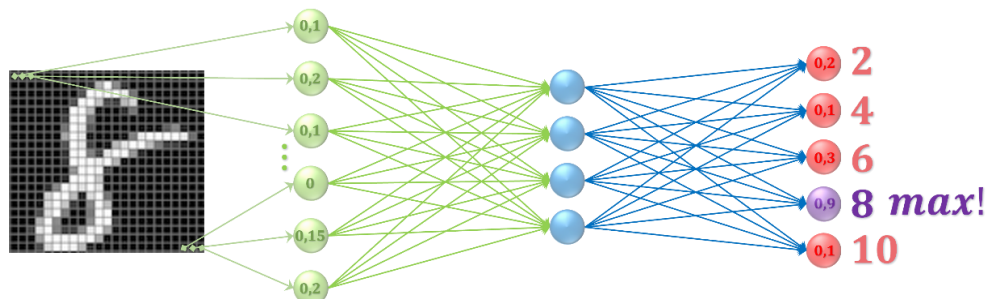
1. Die Werte der Neuronen in Eingabeschicht werden mit den Eingabewerten initialisiert. Das können beispielsweise Helligkeitswerte der einzelnen Punkte des Rasterbilds sein, auf dem eine zu erkennende Ziffer abgebildet ist.



2. Die Werte der darauffolgenden Schicht werden dann anhand der Werte der Neuronen in der Eingabeschicht berechnet, indem jedes einzelne Neuron der nächsten Schicht die Werte seiner Vorgängerneuronen aus der Eingabeschicht ausliest, mit den jeweils für die einzelnen Eingänge hinterlegten Parametern gewichtet und daraus den Ausgabewert ermittelt.



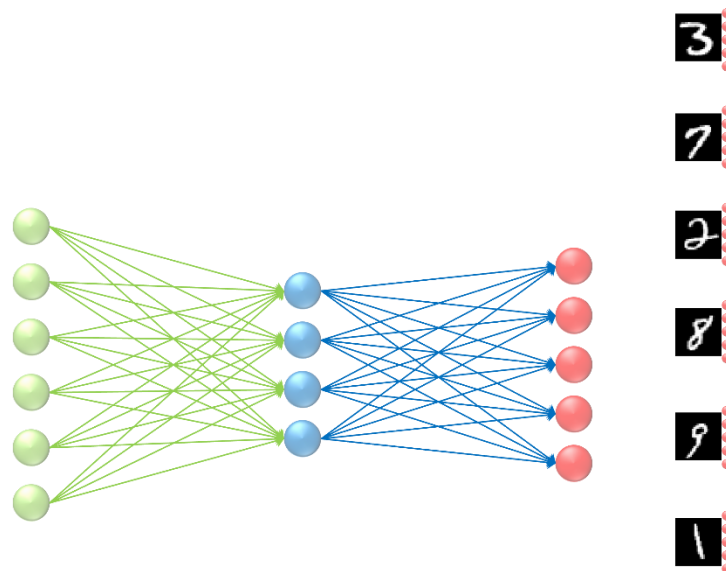
3. Dieser Vorgang wird dann für die nächste Folgeschicht wiederholt, bis man schließlich zur Ausgabeschicht gelangt. Die so berechneten Werte der Neuronen in der Ausgabeschicht bilden nun die berechnete Ausgabe des Netzes. Das könnten beispielsweise 6 Neuronen – je eines die Zahl 2, 4, 6, 8, und 10 – sein, von denen dasjenige, das der Zahl auf dem Eingabebild entspricht, im Idealfall den höchsten Wert hat.



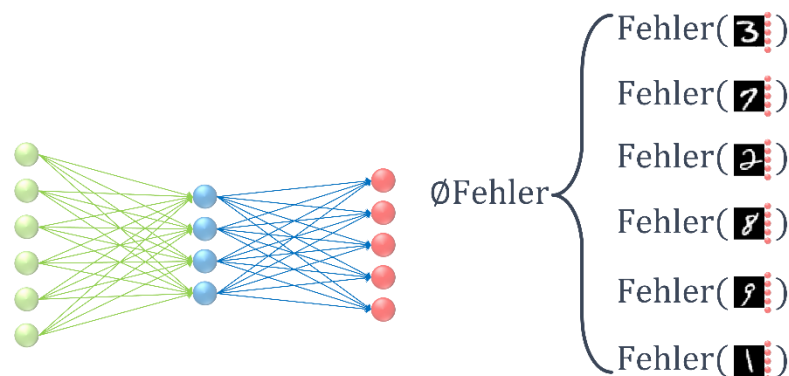
Im Wesentlichen geschieht also in einem solchen Netz nichts anderes, als das fortwährend gewichtete Summen von Eingabewerten der jeweiligen Vorgängerschicht berechnet und als Eingabewerte für die Neuronen der Folgeschicht verwendet werden, bis man schließlich an der Ausgabeschicht ankommt.

Das Potenzial solcher Netze besteht nun aber in der Möglichkeit, die Gewichtungparameter einzelner Neuronen – analog zu den Synapsen natürlicher Netze – zu manipulieren und die Erkennungsleistung des Netzes somit anhand von Lernbeispielen zu trainieren. Dazu geht man in der sogenannten „Trainingsphase“ des Netzes wie folgt vor:

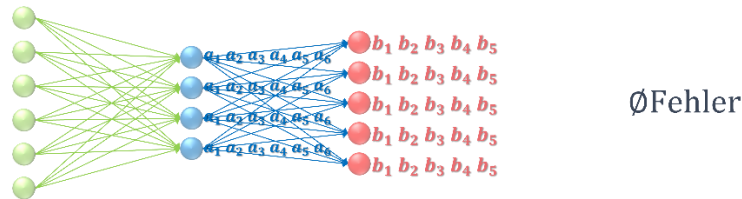
1. Die Gewichtungparameter sämtlicher Neuronen werden mit beliebigen Zufallswerten initialisiert.
2. Anschließend legt man das Eingangssignal eines Lernbeispiels (also beispielsweise das Abbild der Zahl „3“) an und lässt das Netz den Ausgabewert wie oben beschrieben berechnen.
3. Diesen Ausgabewert vergleicht man mit dem bereits bekannten „gewünschten“ Wert zum verwendeten Lernbeispiel (also ein besonders hoher Wert am Ausgabeneuron für „3“ und niedrige Werte an allen anderen Ausgabeneuronen) und merkt sich die festgestellte Abweichung.



4. Das Ganze wird dann für alle Lernbeispiele (also etwa verschiedene Ziffern in verschiedenen Darstellungen) wiederholt. Aus den jeweils gemessenen Abweichungen bildet man dann den Mittelwert (die sogenannte Fehlerfunktion) und manipuliert die Gewichtungparameter der Ausgabeschicht anhand der ersten Ableitung der Fehlerfunktion so, dass die Fehlerfunktion ihrem Minimum zustrebt.



5. Auf Basis dieser Manipulation verändert man die Gewichtungparameter der Vorgängerschicht in analoger Weise, um sich den Werten anzunähern, die für eine bessere Ausgabequalität an der Ausgabeschicht notwendig gewesen wären.



6. Diesen Vorgang wiederholt man dann Schicht für Schicht, bis man an der Eingabeschicht angekommen ist. Das Ganze nennt sich „Backpropagation“ (Rückrechnen).

Diesen gesamten Lernschritt wiederholt man jetzt erneut mit sämtlichen Lernbeispielen und wertet erneut die gemessenen Abweichungen von den gewünschten Ergebnissen aus. Anschließend führt man wieder die eben erläuterte Backpropagation aus, um den Lernerfolg des Netzes weiter zu verbessern. Dieses Vorgehen wird schließlich solange wiederholt, bis die gemessene Abweichung vom gewünschten Ergebnis unterhalb einer vorgegebenen Schwelle liegt. In diesem Moment ist die Trainingsphase des Netzes abgeschlossen.

In der Regel schließt sich an die Trainingsphase eine Testphase an, während derer dem Netz Eingabewerte vorgelegt werden, die nicht in der Trainingsphase verwendet wurden. Je nach Qualität der dabei gemessenen Ausgaben kann das Netz dann entweder in seine „Vorhersagephase“ gehen – also für seine eigentliche Aufgabe eingesetzt werden – oder muss erneut mit einem modifizierten Satz an Lernbeispielen trainiert werden.

Von entscheidender Bedeutung für die Leistungsfähigkeit eines solchen Netzes sind dabei verschiedene Designparameter, wie etwa die Größe und Anzahl der Schichten sowie die geeignete Wahl der Fehlerfunktion. So ist ein zu klein dimensioniertes Netz oft nicht in der Lage, sich die wesentlichen Unterscheidungsmerkmale der Eingabedaten genügend detailliert einzuprägen. Bei einem zu groß dimensionierten Netz besteht hingegen die Gefahr, dass es gar nicht erst nach abstrakten Unterscheidungsmerkmalen sucht, sondern die Lernbeispiele gewissermaßen einfach auswendig lernt. In beiden Fällen wäre ein solches Netz nicht in der Lage allgemeine Eingaben sicher zu erkennen.

Die in diesem Unterabschnitt dargelegte Funktionsweise künstlicher neuronaler Netze dient selbstverständlich nur dem prinzipiellen Verständnis. Netze, wie sie heutzutage in der Praxis eingesetzt werden, enthalten oft noch eine Fülle weiterer Mechanismen und Strukturmerkmale, deren Erläuterung jedoch den Rahmen dieser Ausarbeitung bei Weitem sprengen würden. Nichtsdestotrotz sollte aus den her angestellten Überlegungen deutlich geworden sein, wie ein künstliches neuronales Netz grundsätzlich funktioniert und auf welchen Überlegungen diese Funktionsweise beruht.

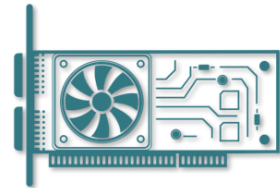
2.3 Historische Entwicklung künstlicher neuronaler Netze

Wie bereits im vorangegangenen Unterabschnitt erwähnt, sind die dort beschriebenen Grundüberlegungen für die Funktionsweise künstlicher neuronaler Netze bereits in den 1940er Jahren erdacht und ihre praktische Nutzbarkeit erstmals in den 1950er Jahren anhand einfacher

Beispiele gezeigt worden. Neben so manchen seinerzeit noch nicht gelösten Grundlagenprobleme erwiesen sich jedoch vor allem zwei Aspekte als vorläufig unüberwindbare Hindernisse für den Einsatz neuronaler Netze in praktisch relevanten Bereichen:

1. Schon für relativ einfache Aufgaben, wie etwa Zeichenerkennung, muss ein Netz eine gehörige Menge Neuronen enthalten, die entsprechend komplex miteinander vernetzt sind. Die dadurch entstehenden kombinatorischen Größenordnungen für die Berechnung der Ausgabewerte und der Backpropagation wachsen daher sehr schnell ins Unermessliche, so dass die Rechenleistung verfügbarer Computersysteme nicht für praktisch nutzbare Systeme ausreicht. So hat GPT-4 (das Sprachmodell, das hinter der jüngsten Version ChatGPT steckt) alleine knapp 100 Milliarden Neuronen und rund 100 Billionen Verbindungen zwischen diesen Neuronen [12].
2. Um etwa komplexere Aufgaben wie Textverständnis trainieren zu können, bedarf es einer entsprechenden Fülle aufbereiteter Beispieldaten – also Texte mit hervorgehobenen Elementen, anhand derer sich ihre Bedeutung beschreiben lässt, so dass man die berechneten Ausgaben des Netzes mit der intendierten Bedeutung des Beispieltexes abgleichen kann. Derartig aufbereitete Lernbeispiele standen lange Zeit schlichtweg nicht in genügender Menge zur Verfügung und ließen sich auch nicht „eben mal schnell“ von einer dafür zusammengestellten menschlichen Arbeitsgruppe erstellen.

Beides änderte sich etwa seit den 2010er Jahren schlagartig. Ab dann standen nämlich Rechnersysteme zur Verfügung, die in besonderem Maße für die Berechnungen optimiert sind, die bei neuronalen Netzen überwiegend anfallen – und zwar in Form von Grafikkarten. Diese hochspezialisierten Rechensysteme sind eigentlich für die rasend schnelle Lösung von Problemen der geometrischen Datenverarbeitung entwickelt worden, wie sie vor allem bei fotorealistischen 3D-Computerspielen anfallen. Zufälligerweise sind diese Berechnungen aber weitgehend mit denen identisch, die im vorigen Unterabschnitt für neuronale Netze beschrieben wurden: in beiden Anwendungsfällen geht um die möglichst effiziente Berechnung von Summen parametrisierter Werte sowie um die Anwendungen einfacher Rechenoperationen aus dem Bereich der linearen Algebra. Damit stand mit einem Mal eine enorme Rechenleistung zur Verfügung, die es endlich ermöglichte, entsprechend komplexe Netzwerke zu designen.



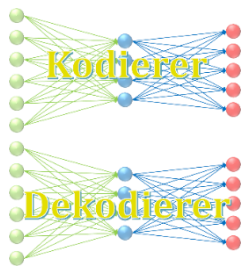
Etwa zur gleichen Zeit begann der Siegeszug des World Wide Webs und vor allem der sozialen Netzwerke. Hauptleistung der großen Suchmaschinen des World Wide Webs war es ja gerade, inhaltliche Merkmale von Webseiten zu erkennen, zu kategorisieren und miteinander in Verbindung zu bringen. Soziale Netzwerke hingegen sorgten zum für eine Unmenge an Text-, Bild- und Videomaterial, das von vernetzten Personen kommentiert und insofern inhaltlich erläutert und verschlagwortet wurde. Zum anderen lieferten die Textkommunikationssysteme (Chat-Apps) enorme Mengen an Vorlagen für typische Dialoge zwischen Menschen. Damit war also auch eine gigantische weltweit produzierte Datenbasis vorhanden, anhand derer man die nun möglichen komplexen Netze sinnvoll trainieren konnte.

Es ist also kein Zufall, dass sich gerade die sogenannten großen Sprachmodelle wie ChatGPT in den Jahren ab ca. 2010 relativ schnell entwickelten. Im folgenden Unterabschnitt

wollen wir daher überblickartig darlegen, welche Besonderheiten diesen großen Sprachmodellen zugrunde liegen und was das für deren Leistungsfähigkeit bedeutet.

2.4 Große Sprachmodelle und ChatGPT

Als Sprachmodelle („Language Models“ bzw. LMs) bezeichnet man Systeme, welche die sprachlich sinnvolle Zusammensetzung von Wortfolgen anhand dafür ermittelter Wahrscheinlichkeiten vorhersagen können. Große Sprachmodelle („Large Language Models“ bzw. LLMs) sind dementsprechend Sprachmodelle mit einer großen Menge an verarbeitetem Wissen über die Wahrscheinlichkeit sprachlich sinnvoller Wortfolgen. Unter den LLMs hat sich in den letzten Jahren das Prinzip des „Transformers“ als besonderes erfolgreich erweisen.



Diese bestehen aus zwei wesentlichen Komponenten: einem „Kodierer“ und einem „Dekodierer“. Der „Kodierer“ hat die Aufgabe, Textsequenzen zu verstehen, während der Dekodierer gewissermaßen die Antwort auf die vom Kodierer verstandene Frage liefern soll (also beispielsweise deren Übersetzung in eine andere Sprache). Bei beiden handelt es sich um eine Zusammenfügung neuronaler Netze, die in bestimmter Weise miteinander interagieren.

Eine besondere Bedeutung kommt dabei dem sogenannten Aufmerksamkeitsmechanismus zu. Er sorgt dafür, dass den einzelnen Elementen von Eingabesequenzen (also etwa den einzelnen Wörtern in einem natürlichsprachlichen Satz) in Abhängigkeit von Bedeutung und Stellung des jeweiligen Worts besondere Aufmerksamkeit zugeordnet wird. Das ist insbesondere wichtig, um Satzteile in den richtigen Kontext zueinander stellen zu können.



Als Beispiel sei der folgende Satz genannt: „Das Tier überquerte das Hindernis nicht, weil es zu müde war“. Ohne tieferes Textverständnis ist es nicht auf Anhieb möglich zu erkennen, ob sich das „es“ in „weil es zu müde war“ auf das Tier oder das Hindernis bezieht. Der Aufmerksamkeitsmechanismus hingegen ordnet den einzelnen Wörtern zum einen aufgrund ihrer gelernten Bedeutung, zum anderen jedoch aufgrund ihrer Stellung im Satz bestimmte Gewichtungen zu, die es ermöglichen, solche Entscheidungen mit hoher Wahrscheinlichkeit richtig zu treffen. Im Kodierer dient der Aufmerksamkeitsmechanismus dabei vor allem der kontextabhängigen Interpretation der einzelnen Eingabebestandteile, während er im Dekodierer der kontextabhängigen Zusammensetzung der Ausgabebestandteile dient.

Die für den Aufmerksamkeitsmechanismus erforderlichen Erkenntnisse über die Struktur von Sprache, die Bedeutung von Wörtern und die Auswirkungen verschiedener Satzstellungen werden in der Regel im Rahmen einer automatisierten Analyse sehr großer Mengen exemplarischer Textvorlagen gewonnen, mit deren Hilfe ein entsprechend konfiguriertes neuronales Netzwerk trainiert wird. Diesen Vorgang nennt man „Pre-Training“. An diese Pre-Training-

Phase schließt sich dann das sogenannte Finetuning an, während dessen der Transformer anhand von Menschen annotierter Texte auf spezielle Aufgaben hin (Sprachübersetzung, Textzusammenfassung oder das Beantworten von Fragen) optimiert wird. Mit diesen Hilfsmitteln versetzt man den Transformer insbesondere in die Lage, selbständig Texte zu generieren, die denjenigen aus den Trainingsdaten strukturell möglichst ähnlich sind – eine Fähigkeit, die gemeinhin als „generativ“ bezeichnet wird.

Einen derartig auf Pre-Training basierten generativen Transformer hat die Firma OpenAI erstmals im Jahre 2017 unter dem Kürzel „GPT“ („Generative Pre-trained Transformer“) auf den Markt gebracht. Mittlerweile ist bereits die vierte Version davon unter dem Akronym „GPT-4“ auf dem Markt. Das auf GPT-3.5 und GPT-4 basierte Konversationssystem wird unter dem Namen „ChatGPT“ vermarktet. GPT-4 ist auf Beispieltextrn mit einem Gesamtvolumen von 45TB [14] trainiert worden. Unterstellt man, dass ein Taschenbuch rund 200 Seiten à 1.700 Zeichen – mit hin also 340.000 Zeichen (also rund 330KB) hat, so entspricht das rund 150 Millionen Taschenbüchern. Denkt man sich ein Regal von einem Meter Breite und zwei Metern Höhe, das also zehn Böden à 100 Taschenbücher – mit hin also 1.000 Taschenbücher – fasst, so könnte man eine 150km lange Reihe (was etwa der Entfernung Hamburg/Hannover entspricht) aus solchen Regalen bilden, um die Menge an Text zu lagern, mit der die GPT-4 trainiert wurde.

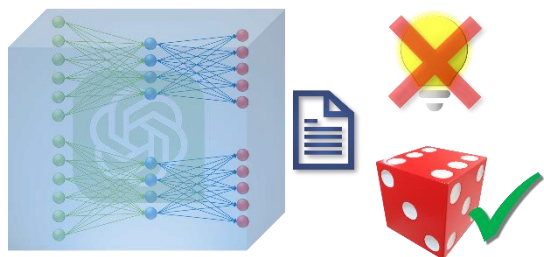


Eine weitere Besonderheit der Transformer-Architektur besteht darin, dass sämtliche Bestandteile eines zusammenhängenden Textes in weiten Teilen unabhängig voneinander in Kodierer und Dekodierer verarbeitet werden können. Damit lässt sich eine hochgradige Parallelisierung des Verarbeitungsprozesses erreichen, durch welche derartige System in die Lage versetzt werden, auch sehr große Texte mit hoher Geschwindigkeit zu analysieren bzw. zusammenzustellen.

Zusammenfassend kann man sagen, dass es sich bei Systemen nach Bauart von ChatGPT um ein sehr komplexes Zusammenspiel sehr großer neuronaler Netze handelt, die mit einer gigantischen Datenbasis trainiert wurden und insbesondere mit Hilfe des Aufmerksamkeitsmechanismus in die Lage versetzt werden, aus einer umfassenden Analyse der Trainingsdaten heraus auf verschiedenste Fragestellungen hin Texte zu generieren, die mit einer hohen Wahrscheinlichkeit zur jeweiligen Fragestellung passen.

3 Stärken und Grenzen von ChatGPT & Co.

Wie im vorangegangenen Abschnitt herausgearbeitet, handelt es sich bei Systemen nach Art von ChatGPT um sehr umfassend trainierte neuronale Netze, die mit Hilfe ausgefeilter Analysen von Sprachstruktur und Bedeutungszusammenhängen von Wörtern auf konkrete Fragestellungen hin Texte generieren, die mit einer hohen Wahrscheinlichkeit als Antwort auf solche Fragestellungen passen. Davon abzugrenzen sind indessen Systeme, die tatsächlich die Inhalte der trainierten Texte verstehen und daraus mit Hilfe eines intelligenten Inferenzmechanismus Schlüsse ziehen können. ChatGPT & Co. bilden



die Sätze ihrer Ausgabertexte letzten Endes ausschließlich auf Basis von Wahrscheinlichkeitsverteilungen. Die Frage, ob man mit solchen Systemen das eigentliche Sprachverstehen umfassend simulieren kann, ist derzeit streitig [4]. ChatGPT selbst generiert jedenfalls in vielen Fällen außerordentlich wohlklingende Texte, deren Faktentreue allerdings nicht gewährleistet werden kann und die schon gar nicht von einem echten inhaltlichen Verständnis der verarbeiteten Texte zeugt.

Die Frage, ob man mit solchen Systemen das eigentliche Sprachverstehen umfassend simulieren kann, ist derzeit streitig [4]. ChatGPT selbst generiert jedenfalls in vielen Fällen außerordentlich wohlklingende Texte, deren Faktentreue allerdings nicht gewährleistet werden kann und die schon gar nicht von einem echten inhaltlichen Verständnis der verarbeiteten Texte zeugt.

Das mag gut und gerne reichen, um etwa das Grundgerüst eines Aufsatzes oder eine alltägliche E-Mail-Nachricht zu generieren, gleichwohl man die darin postulierten Fakten immer nochmals einer kritischen Überprüfung unterziehen sollte. Wenn es jedoch darum geht, den semantischen Sinn eines Textes wirklich zu verstehen, kann man sich auf ChatGPT keinesfalls verlassen. Ein schönes Beispiel hat der Autor dieser Ausarbeitung selbst mit ChatGPT (GPT-3.5) generiert – und zwar in Bezug auf ein Forschungsgebiet, in dem er sich zu akademischen Zeiten betätigt hat.



Konkret geht es um die sogenannte intuitionistische Logik. Während die klassische Logik sich mit der Frage nach der Wahrheit von Aussagen beschäftigt, geht es in der intuitionistischen Logik um die Frage, ob eine Aussage beweisbar ist. Das führt insbesondere dazu, dass das klassische Prinzip „tertium non datur“ („das ausgeschlossene Dritte“) in der intuitionistischen Logik nicht allgemein gilt. Während also in der klassischen Logik für eine beliebige Aussage „A“ die Behauptung „A oder nicht A“ immer wahr ist (Beispiel: „das Wetter ändert sich oder es bleibt, wie es ist“ stimmt immer, denn es kann niemals sein, dass das Wetter sich weder ändert noch so bleibt, wie es ist), kann es in der intuitionistischen Logik durchaus möglich sein, dass man weder „A“ noch sein Gegenteil beweisen kann. Wenn man also etwa gerade aufgewacht ist und aus dem Fenster schaut, kann man nicht auf Anhieb sagen, ob das Wetter sich gerade ändert oder so bleibt, wie es ist. Dazu muss man es erst eine Weile lang beobachten. Es ist also durchaus möglich, dass man zu einem gegebenen Zeitpunkt weder beweisen kann, dass sich das Wetter gerade ändert, noch, dass es gerade bleibt, wie es ist.

Hierzu hat der Autor dieser Ausarbeitung folgenden Dialog mit ChatGPT geführt (dabei ist mit „ $\neg C$ “ die Negation einer Aussage „C“ gemeint – gelesen als „nicht C“):

Frage: „Ist ‚C oder $\neg C$ ‘ intuitionistisch gültig?“

ChatGPT: „Ja, die Aussage ‚C oder $\neg C$ ‘ ist in der intuitionistischen Logik gültig. Anders als in der klassischen Logik, die das Prinzip des ausgeschlossenen Dritten (Law of

Excluded Middle, LEM) akzeptiert, wird in der intuitionistischen Logik das Prinzip des ausgeschlossenen Dritten abgelehnt.“

ChatGPT behauptet also zunächst fälschlicher Weise, dass „C oder nicht C“ intuitionistisch **gültig** sei, erklärt dies aber damit, dass das Prinzip des ausgeschlossenen Dritten in der intuitionistischen Logik **abgelehnt** werde, was ja auch stimmt. Die komplett falsche Behauptung von ChatGPT begründet es also mit ihrem genauen Gegenteil (das immerhin zutreffend ist)!

Interessanter Weise beantwortet ChatGPT dieselbe Frage in nur leicht modifizierter Formulierung absolut korrekt und begründet die Antwort sehr akkurat (wenngleich etwas redundant und mit einer gewissen schwülstigen Tonalität):

Frage: „Ist C oder $\neg C$ in intuitionistischer Logik gültig?“

ChatGPT: „In der intuitionistischen Logik ist die Aussage ‚C oder $\neg C$ ‘ nicht immer gültig, es sei denn, es liegt ein konstruktiver Beweis für C oder $\neg C$ vor. Die intuitionistische Logik basiert auf einer anderen Sichtweise auf Logik und Wahrheit als die klassische Logik. In der intuitionistischen Logik wird die Wahrheitswertigkeit von Aussagen oft in Bezug auf die Verfügbarkeit konstruktiver Beweise betrachtet.

Wenn ein konstruktiver Beweis für C oder $\neg C$ vorliegt, ist die Aussage gültig. Dies bedeutet, dass entweder ein Beweis für C oder ein Beweis für $\neg C$ präsentiert werden muss, damit die Aussage als wahr in der intuitionistischen Logik betrachtet wird.

Wenn jedoch kein konstruktiver Beweis für C oder $\neg C$ vorliegt, bleibt die Wahrheitswertigkeit dieser Aussage in der intuitionistischen Logik unklar, und sie wird nicht als automatisch gültig angesehen. Dies steht im Einklang mit dem intuitionistischen Prinzip, dass die Existenz eines Beweises die Grundlage für die Wahrheit einer Aussage ist.“

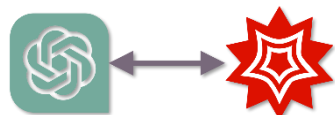
An diesen Beispielen ist gut erkennbar, dass ChatGPT eben *gerade nicht* wirklich versteht, was in den Trainingstexten niedergelegt ist. Stattdessen generiert es zu einer gegebenen Fragestellung zugegebenermaßen sehr eloquent formulierte Texte auf Basis von Wahrscheinlichkeitsverteilungen, die es aus der Analyse der Trainingstexte hergeleitet hat. In vielen Fällen passt das hervorragend als Antwort auf die gestellte Frage, aber verlassen kann man sich darauf keinesfalls. In gewisser Weise kann man es auch so ausdrücken, dass Sprachmodelle nach Art von GPT im Moment letztlich nichts anderes sind als mehr oder weniger überzeugende digitale Hochstapler [10].

4 Die Zukunft generativer Systeme

Die im vorangegangenen Abschnitt entwickelte Sichtweise, dass es sich bei Systemen nach Art von GPT letztlich um nichts weiter als mehr oder minder überzeugende digitale Hochstapler handelt, muss allerdings nicht heißen, dass es in Zukunft so bleiben wird. Die Entwicklung der generativen Systeme ist in den letzten Jahren extrem schnell vorangeschritten. Es ist also durchaus wahrscheinlich, dass sich die Systeme sehr schnell sehr stark verbessern werden. Letztlich bedeutet das aber bis auf Weiteres nur, dass sich die Wahrscheinlichkeiten für falsche Aussagen in Zukunft deutlich verringern werden. Ob generative Systeme eines Tages so gut

hochstapeln können, dass die von ihnen erzeugten Resultate sich in der Praxis zumindest nicht mehr von denjenigen einschlägiger Fachleute unterscheiden lassen, ist gegenwärtig noch nicht genau vorhersagbar.

Gut möglich ist indessen, dass sich die auf neuronalen Netzen basierenden Techniken mit herkömmlichen Techniken zur Automatisierung von Schlussfolgerungen aus bestehenden Fakten kombinieren lassen, um zumindest in abgegrenzten Teilbereichen wesentlich belastbarere Ergebnisse erzielen zu können. Schon jetzt gibt es beispielsweise ein [Plugin für ChatGPT](#), das auf die Wolfram Language von Wolfram Alpha zugreifen und somit die Qualität der von



ChatGPT erzeugten Aussagen zumindest für Problemstellungen aus der von Wolfram Alpha abgedeckten Domäne von Mathematik und Naturwissenschaft erheblich verbessern kann.

Darüber hinaus gibt es schon jetzt eine Fülle generativer Systeme, die uns viel Arbeit abnehmen können, wenn man ihre Grenzen einschätzen und entsprechend mit ihnen umgehen kann. So kann das ebenfalls von openAI entwickelte Bildgenerierungssystem „DALL-E“ erstaunlich gelungene [Bilder aus kurzen Textbeschreibungen](#) erzeugen. Auch gibt es bereits sehr nützliche Systeme zur Erzeugung von Präsentationen aus vorgegebenen textlichen Beschreibungen, wie etwa „[presentations.ai](#)“. Ähnliche Systeme gibt es natürlich auch für die Erzeugung von Musik und Videos auf Basis textlicher Beschreibungen.

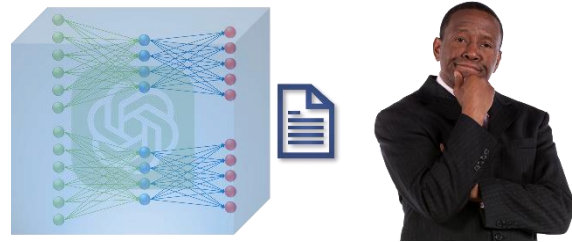


Ein äußerst kritischer Nebeneffekt dieser enormen Leistungsfähigkeit generativer Systeme sind sogenannte „Deepfakes“ (Fälschungen, die mit Hilfe tiefgehend lernender Systeme erstellt werden). So können derartige Systeme für Bild-, Ton- und Videoerzeugung schon heute nahezu täuschend echte Bilder, Tonaufnahmen bzw. Videos erzeugen, auf denen reale Personen zu hören bzw. zu sehen sind, deren hörbare Aussagen bzw. sichtbare Handlungen

jedoch künstlich von diesen Systemen erzeugt wurden. Derartige Deepfakes können im besten Falle für satirische Zwecke im schlimmsten Falle jedoch für hochgradig kriminelle Zwecke (Betrug, Verleumdung, Vertuschung) eingesetzt werden. Es liegt auf der Hand, dass hierin ein nicht zu unterschätzendes Risiko im Zusammenhang mit dem Einsatz generativer KI besteht, deren Verwendung insoweit einem dringenden Regulierungsbedarf unterliegt.

Trotz dieses Missbrauchspotenzials deutet dennoch alles darauf hin, dass uns schon in aller nächster Zukunft generative Systeme dieser Art eine Menge kreativer Basisarbeit abnehmen werden, sofern wir lernen, richtig mit ihnen umzugehen. Ob das, wie vielfach heraufbeschworen, zum baldigen Aussterben ganzer Berufszweige führen wird, darf bezweifelt werden. Voraussetzung dafür wäre jedenfalls, dass die Systeme – wie oben dargelegt – sehr viel fakten-treuer und verlässlichere Ergebnisse produzieren als bisher, wobei es zur Beurteilung dieses Umstands überhaupt erst anerkannte Qualitätsmaßstäbe und -tests geben muss.

Bis dahin ist kritischer Sachverstand für die Nutzung solcher Systeme unerlässlich. Diesen vorausgesetzt kann man jedoch durchaus schon jetzt sehr produktiv mit solchen Systemen arbeiten. Das mag uns für den Moment überfordern und gewisse Ängste vor einer Bevormundung durch solche Systeme schüren.



Letztlich werden sie aber früher oder später so selbstverständlich in unsere Arbeitsabläufe integriert werden, wie heute schon die vielen informationstechnologischen Helferlein wie Textverarbeitung, CAD, 3D-Drucker und vieles mehr, von denen wir uns höchst gerne und dankbar die vielen un kreativen Arbeiten abnehmen lassen, die dereinst von eigens für solche Tätigkeiten qualifizierten Fachleuten (also etwa Schreibkräfte, technische Zeichner, Modellbauer) erledigt werden mussten, auch wenn sich so mancher Nostalgiker die Zeiten zurückwünschen mag, in denen es diese Berufsbilder noch gab.



Der aktuelle KI-Hype lehrt uns aber auch, dass wir uns schon jetzt auf den wohl nicht mehr allzu fernen nächsten großen informationstechnologischen Quantensprung einstellen müssen, aus dem Systeme hervorgehen, die uns kognitiv ebenbürtig bzw. wohl auch schon sehr bald überlegen sein werden. Es bleibt zu hoffen, dass wir bis dahin gelernt haben, verantwortungsvoll mit solchen Entwicklungen umzugehen und uns weltweit gültige Regeln zu verordnen, die eine unkontrollierte Verselbständigung solcher uns überlegener Systeme wirksam verhindern.

5 Quellen

1. „Aufmerksamkeit reicht“, c't Magazin 2022/11, S. 136
2. „Der universelle Texter“, c't Magazin 2023/05 S. 56
3. „Neural Networks Without the Math“, Joyously Aware Media, 2018
4. „Schneller als gedacht“, c't Magazin 2023/09, S. 126
5. <http://jalanmar.github.io/illustrated-transformer/>
6. <https://de.wikipedia.org/wiki/Deepfake>
7. https://de.wikipedia.org/wiki/K%C3%BCnstliches_neuronales_Netz
8. https://en.wikipedia.org/wiki/Language_model
9. https://en.wikipedia.org/wiki/Large_language_model
10. <https://encord.com/blog/generative-ai-and-gpt4-is-overrated-long-live-old-school-ai/>
11. <https://encord.com/glossary/gpt-definition/#:~:text=GPT%2C%20or%20Generative%20Pre%2Dtrained,closely%20resemble%20human%2Dwritten%20text.>
12. <https://medium.com/@fenjiro/chatgpt-gpt-4-how-it-works-10b33fb3f12b>

13. <https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile#:~:text=The%20latest%20version%20of%20the,of%20273%20set%20by%20Arizona>.
14. <https://www.projectpro.io/article/gpt3-vs-gpt4/816#:~:text=Dataset%20Size,45%20gigabytes%20of%20training%20data>.
15. <https://zapier.com/blog/how-does-chatgpt-work/>
16. <https://youtu.be/zxQyTK8quyY?si=Y4CdXtcfuMNPu4X9>
17. <https://youtu.be/kWLed8o5M2Y?si=FefwSxny7OkFr3uY>