Künstliche Intelligenz: Künstlich? – Ja. Intelligent? – Nein.

Shelia Guberman, Cupertino/CA, USA[1]

Egal, wie oft Sie „Halva, Halva" sagen,
Die Süße im Mund werden Sie nicht spüren.

Khoja Nasreddin

Redaktionelle Vorbemerkung

In den letzten 10 Jahren ist die Popularität der künstlichen Intelligenz (KI) dank der Entwicklung neuer Versionen der neuronalen Netze (NN) - Bilderkennung und große Sprachmodelle - extrem schnell gewachsen. Die Errungenschaften sind real und decken viele Bereiche der Wirtschaft, des Alltags und verschiedener menschlicher Aktivitäten ab. Der gestaltpsychologisch orientierte Autor ist Atomphysiker, Entwickler eines der ersten Texterkennungs-Programme, Pionier in vielen Bereichen von Erkennungs-Software und zugleich ein überaus sachkundiger Kenner dieser Technologien. Er ist auch ein aktiver Benutzer von ChatGPT. Der Zweck seines Artikels ist es, die inhärenten Mängel dieses Tools herauszuarbeiten, die großen Schaden anrichten können.

Dass ein solcher Beitrag in einer Psychotherapie-Zeitschrift erscheint, mag auf den ersten Blick erstaunlich sein. Der letzte Satz des Beitrags mag deutlich machen, warum er von so allgemeiner Bedeutung ist – auch für den psychotherapeutischen Bereich: „Heute wird das tragische Ereignis der Ersetzung des Denkens durch Auswendiglernen von der KI-Gemeinschaft verkündet und von den Massenmedien als großer Schritt zur Intelligenz dargestellt."
Die Terminologie und die Darstellung der technologischen Herausforderungen der KI in diesem Beitrag wird manchen Leserinnen und Lesern vielleicht sehr ungewohnt sein – wir denken jedoch, dass es sich gerade auch für Psychotherapeutinnen lohnt, sich mit den Leitgedanken und der Logik dieser Entwicklungen auseinanderzusetzen.

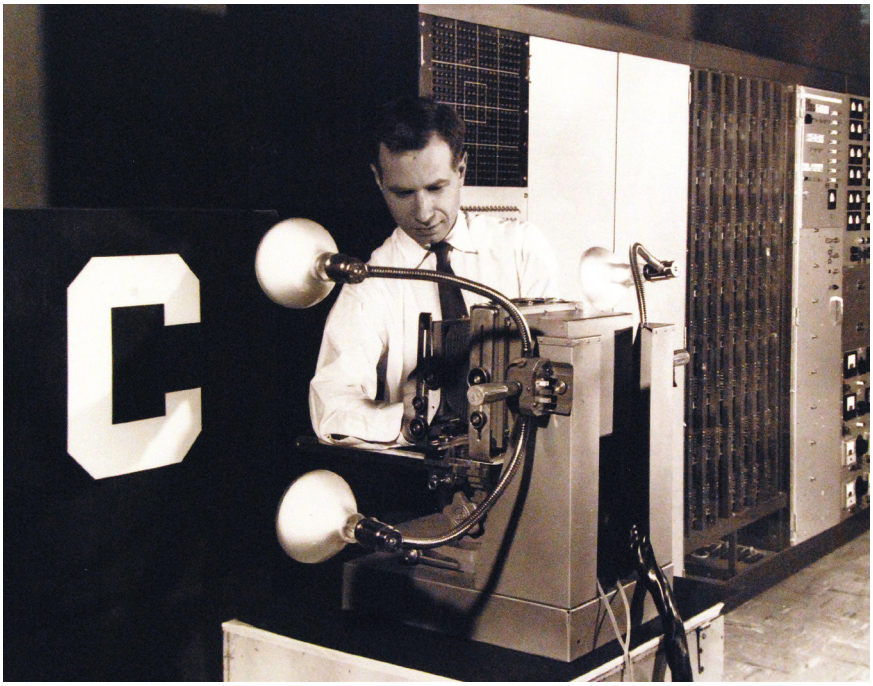
Eine kurze Geschichte der Neuronalen Netze

Die Geschichte der KI wurde bereits mehrfach geschrieben. Immer beginnt diese Geschichte mit dem Neuronenmodell von McCulloch und Pitts zur Nachbildung realer Vorgänge in neuronalen Strukturen (McCulloch & Pitts 1943) und der philosophischen Abhandlung von Turing „Computing Machinery and Intelligence" (Turing 1950). Wie blickten diese Pioniere in die Zukunft? McCulloch und Pitts waren sehr optimistisch und stellten der Psychiatrie, der Neurophysiologie und der Computerbiologie große Möglichkeiten in Aussicht. Zum Beispiel: „Für den Psychiater bedeutet das, dass in solchen Systemen der ‚Geist' nicht mehr ‚gespenstischer als ein Gespenst' ist. Stattdessen kann die gestörte Psyche ohne Verlust an Umfang oder Strenge mit den wissenschaftlichen Begriffen der Neurophysiologie verstanden werden" (McCulloch & Pitts 1943; übers.). Turing näherte sich dem Problem auf eine grundsätzliche Weise, indem er die grundlegenden Begriffe der *Maschine* und des *Denkens* definierte. Dabei ist zu betonen, dass keiner von ihnen behauptete, das Ziel sei die Schaffung einer intelligenten Maschine – einer Maschine, die denken kann. McCulloch und Pitts wollten vielmehr die Netze künstlicher Neuronen nutzen, um die Funktionsweise des menschlichen Gehirns zu verstehen, und Turing erklärte, dass es ihm um die *Nachahmung* des Denkens ginge.

Der nächste Schritt war das 1957 von Rosenblatt vorgeschlagene *Perceptron* (Rosenblatt 1957). Dabei handelte es sich um ein analoges elektrisches Gerät, das die Idee des Lernens umzusetzen versuchte – ein Konzept, das diesen Bereich der KI bis heute beherrscht. Das Perceptron war für die Bilderkennung gedacht, und die ersten Objekte, die erkannt wurden, waren Buchstaben und Ziffern. Dieses Thema zieht sich durch die gesamte Geschichte der KI. Die Erörterung dieser Entwicklungslinie wird es uns schließlich ermöglichen, das Wesen der modernsten Produkte neuronaler Netze zu verstehen – die Großen Sprachmodelle (Large Linguistic Models = LLM).

Schon sehr bald nach dem Erscheinen von Rosenblatts Arbeit wurde das von ihm vorgeschlagene

1 Leicht gekürzte deutschsprachige Fassung des Beitrags „Artificial Intelligence vs Intelligence", https://www.academia.edu/121505671/ – Rohübersetzung durch DeepL, überarbeitet von Gerhard Stemberger.



Fotocredit: © Wikimedia - Rosenblatt and the perceptron

Lernverfahren verallgemeinert. Von Anfang an war klar, dass das Perceptron die Buchstaben nicht verallgemeinern konnte: Trainierte es ein bestimmtes „a“, konnte es das gleiche „a“ in einer etwas nach unten verschobenen Stellung nicht erkennen. Dann wurden die Buchstaben als Vektoren präsentiert (z.B. Anzahl der Löcher, der Enden, der Schnittpunkte der Bahnen usw.). Die Trainingsdaten bestanden aus einem Dutzend Beispielen für jeden Buchstaben. Es wurden verschiedene Algorithmen verwendet, aber die Ergebnisse (die Anzahl der Fehler) bewegten sich nach wie vor nicht in einem akzeptablen Bereich und die Gründe dafür waren klar: die ungewöhnliche Form oder Größe oder Lage oder Neigung der Abbildungen der Buchstaben und Ziffern. Die Lösung lag auf der Hand: Die Vielfalt der Buchstaben und Ziffern im Trainingssatz musste erhöht werden. In den 1970er-Jahren erreichte die Anzahl der Beispiele für jeden Buchstaben im Training 200, aber die Ergebnisse waren noch immer nicht gut.

In den 1980ern kehrte das Perceptron auf die Bühne zurück, nachdem es in ein mehrschichtiges Gerät mit der Fähigkeit zum schnellen Vorlauf umgebaut worden war. Das einschichtige Perceptron konnte nur linear trennbare Muster lernen. Die versteckten Schichten projizieren den Eingaberaum in einen Raum höherer Dimension, in dem die beiden Klassen von Punkten, die im Eingaberaum untrennbar waren, linear sicher getrennt werden können. Um die trennende Hyperebene zu finden, wurde die Gradient Slant Methode zur Minimierung der Verlustfunktion angewandt. Für die Erkennung war das gleichbedeutend mit der Erfindung der Wasserstoffbombe – es versprach eine Lösung für jedes Problem. Allerdings warnten die Wissenschaftler, die die Bombe entwickelten, vor den Gefahren der Waffe, während die Adepten der NN (der künstlichen *Neuronalen Netzwerke*) den Mund hielten.

Wie das Superpower-Tool eingesetzt wurde, werden wir am Beispiel der Erkennung handgeschriebener Ziffern durch die NN demonstrieren. Dies ist eines der am wenigsten komplizierten Probleme und es gibt einen großen Datensatz MNIST (LeCun et al. 1998), der zum Testen von NN mit unterschiedlicher Architektur verwendet wurde. Allen gemeinsam ist die Funktion der Ähnlichkeit zweier Bilder – das vom Perceptron geerbte Maß der Überlappung. Dieses Maß geht davon aus, dass die Ziffern in allen Bildern der Trainings- und der Testmenge gleich groß, gleich dick und an der gleichen Position sein müssen. Schon diese Einschränkung zeigt, wie weit die NN vom menschlichen Verstand entfernt sind. Folglich sind eine Menge künstlicher Anpassungen erforderlich und trotz allem wird das Ziel der KI nicht erreicht, etwas wie Intelligenz zu demonstrieren.

Im Jahr 1990 lag der Erkennungsgrad der NNs für handgeschriebene Ziffern bei 85 %. Es gab zwei Methoden zur Verbesserung: Vergrößerung der Trainingsmenge und Vergrößerung der Dimension des Raums, in dem die Klassen getrennt werden müssen. Der gegenwärtige Stand der Technik wurde in einer ausgezeichneten Übersicht von LeCun et al. beschrieben (LeCun 1990). Eine Analyse des Stands der Technik bei der Ziffern-Erkennung zeigte im Jahr 2012, dass verschiedene Erkennungsfunktionen (NNs unterschiedlicher Architektur, SVM und sogar „K-nearest“) bei den meisten populären MNIST-Daten ein Fehlerniveau von weniger als 1 % erreichen konnten. Das „K-nearest“-Ergebnis, das mit K = 1 erzielt wurde, zeigt an, dass jede erkannte Ziffer im Test ein extrem ähnliches Bild in den Testdaten hat („extrem ähnlich“ bedeutet hier, dass sich zwei Bilder fast vollständig überlappen).

Fehleranalyse

In den experimentellen Wissenschaften (und auch im Leben) ist die Fehleranalyse die entscheidende Voraussetzung für das Vorankommen. Lassen Sie uns das also tun. Hier sind einige Beispiele für falsch erkannte Ziffern, die von einer der erfolgreichen Versionen von NN gemacht wurden (Fehlerquote 0,82%).

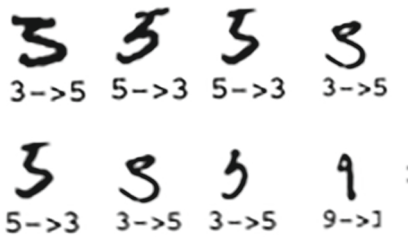


Abb. 1. Zahlen unter den Bildern: Links - tatsächliche Bedeutung, rechts - erkannt als. Im ersten Fall also: 3 erkannt als 5.

(1) Die Gründe für die Fehler beim Erkennen der Ziffern-Bilder in der Abbildung 1 sind offensichtlich: Die ersten fünf Bilder in der Abbildung zeigen die Fehler bei der Unterscheidung von „3“ und „5“. Beide Ziffern bestehen aus zwei Teilen: dem horizontalen Strich (oder Bogen) oben und der S-ähnlichen Figur darunter. Weiters ist die S-Figur (oft in einem einzelnen Strich geschrieben) mit dem rechten Ende des oberen Balkens in der „3“ und mit dem linken Ende in der „5“ verbunden. Die „3“ wird geschrieben, ohne den Stift vom Papier abzuheben. Der Mensch sieht, dass das zweite und dritte Bild in Abb. 1 mit einer Unterbrechung gezeichnet sind, folglich werden beide als „5“ und das erste als „3“ interpretiert.

2) Das NN-Urteil ist leicht zu verstehen: Der zweite Teil der beiden Ziffern sieht sehr ähnlich aus und nimmt den größten Teil der Ziffernfläche ein, was bedeutet, dass er den größten Teil des Ähnlichkeitswertes ausmacht. Dementsprechend kann die Entscheidung nur anhand der Ähnlichkeiten des S-Teils getroffen werden. Die fünfte Ziffer kann von einem Menschen weder als 3 noch als 5 erkannt werden. Das NN hat die Entscheidung getroffen, die Ziffer mit dem ähnlichsten S-Teil zu finden. Obwohl also Tausende von Ziffern „3“ und „5“ richtig erkannt wurden, wird die Erkennungsfunktion den Turing-Test nicht bestehen – ein Mensch hingegen würde solche Antworten nicht geben.

Ein ähnliches Verhalten von NN kann in den nächsten Beispielen beobachtet werden.

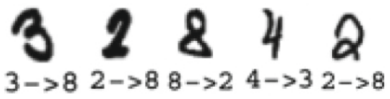


Abb.2: Zahlen unter den Bildern: Links - tatsächliche Bedeutung, rechts - erkannt als. Im ersten Fall also: 3 erkannt als 8.

Hier ist ein weiteres Beispiel für Bilder, die für Menschen zu 100 % erkennbar sind und von der NN fälschlicherweise als 8, 8, 2, 3 und 8 erkannt wurden .

Es ist leicht zu verstehen, dass, wenn der MNIST-Trainingsdatensatz 6000 Ziffern für jede der 10 Ziffern enthält, 1) Bilder für „8“ vorkommen, die die erste, dritte und fünfte Ziffer in unserem Beispiel zu mehr als 80% überlappen, und es 2) vorkommt, dass es in den Trainingsdaten keine Bilder von 2 und 3 gibt, die mit diesen drei Bildern zu mehr als 90% übereinstimmen. Dies zeigt einmal mehr, dass das Maß der Ähnlichkeit (Überschneidung) für die Aufgabe nicht geeignet ist. Es deutet darauf hin, dass das NN nicht generalisieren konnte. Wir werden die Ursache dafür später diskutieren.

(2) Datenanreicherung bedeutet, dass jedes Bild der Trainingsdaten verschiedenen Arten von Verzerrungen (Verschieben, Kippen, Drehen und sogenannte elastische Verzerrungen) unterzogen wird und diese dem Trainingssatz hinzugefügt werden. Dies wurde gemacht, weil das NN keine Ziffer aus dem Trainingssatz korrekt erkennen kann, wenn diese auch nur um zwei Pixel nach rechts verschoben ist, da das Maß der Ähnlichkeit der Grad der Überlappung ist. Dies bedeutet, dass die Generalisierung gleich Null ist.

(3) Vor etwa 20 Jahren wurde entdeckt, dass Bilder, die vom NN korrekt klassifiziert wurden (Abb. 3), bereits nach winzigen, für das Auge unsichtbaren Störungen (Abb. 4) für das NN unerkennbar werden (sogenannte gegnerische Bilder).

Abb. 3

Abb. 4

Das folgende Beispiel eines solchen gegnerischen Bildes kann als eine ernste Warnung dienen:

Abb. 5. Das Bild „STOP“ mit hinzugefügtem Rauschen (links) wurde als das rechte Bild „erkannt“.

Abb. 6. Falsch klassifizierte natürliche Bilder mit hinzugefügtem Rauschen in der rechten unteren Ecke.

In den Schlussfolgerungen des von den Schöpfern des GPT-4 [OpenAI 2023] herausgegebenen Berichts wurden die Aufgaben von morgen aufgelistet. Die wichtigste davon ist der Kampf gegen sogenannte gegnerische Angriffe. Darauf im Einzelnen einzugehen, würde an dieser Stelle aber zu weit führen.[2]

Die Wurzeln der Fehler

Die Analyse der NN-Fehler zeigt, dass diese NN nur solche neuen Ziffern erkennen, die nahezu ideale Kopien einer der Ziffern aus dem Trainingssatz sind – sie beherrschen also keine Verallgemeinerung und verfügen über keine Intelligenz. Das oben beschriebene Verhalten von NN ist für Menschen unmöglich, daher kann nicht behauptet werden, dass NN als Modelle der Funktionsweise des menschlichen Gehirns gelten können.

Welche Fehler wurden bei der Entwicklung der NN gemacht?

1. Die Vorstellung, dass das Gehirn aus Neuronen aufgebaut ist, ist ebenso falsch wie die Vorstellung, dass das Gehirn aus Zellen besteht. Aus philosophischer und psychologischer Sicht ist das Gehirn ein Ganzes. Ein Ganzes ist nach der Definition durch Teile und deren Beziehungen im Ganzen definiert. Wir können das Gehirn nicht beschreiben und analysieren, ohne die Teile zu kennen. Die komplexen Ganzheiten (einschließlich des Gehirns) sind Hierarchien von Teilen. Wir sind uns nicht sicher, ob wir die Hierarchie der Gehirnteile verstehen, aber wir sind uns zu 100 % sicher, dass Neuronen nicht die erste Ebene der Gehirnhierarchie sind, d. h. dass das Gehirn aus Neuronen besteht. Es war also eine schlechte Idee, Modelle der Gehirnfunktion auf der Grundlage des Perceptrons zu erstellen.

2. Ein weiterer Grundgedanke der KI war der Glaube, dass das neuronale Netz kein Anfangswissen benötigt – alles, was benötigt wird, würde während des Lernprozesses extrahiert. Das trifft nicht zu, denn wir müssen sicher sein, dass der Input, die Beschreibung der analysierten Objekte, eine angemessene Information enthält. So beruhte beispielsweise die Ölexploration jahrzehntelang auf der Theorie des organischen Ölursprungs und sie war nahezu zufällig (oder sogar noch schlimmer). Der Einsatz von KI war erfolglos, weil die Theorie falsch war. Erst nachdem die Theorie des nicht-organischen Ölursprungs akzeptiert wurde und dementsprechend der Untersuchungsgegenstand und seine Beschreibung geändert wurden, verbesserte der Einsatz derselben KI-Tools die Erfolgsquote der Ölexploration dramatisch (Guberman 2007).

3. Die NN übernehmen die für das Perceptron getroffene Wahl, die zu verarbeitenden Ziffern als Bitmaps darzustellen. Dies stand im Einklang mit den allgemein anerkannten neuropsychologischen Annahmen aus der Mitte des 20. Jahrhunderts über die Bildverarbeitung in unserem Gehirn. Später wurde festgestellt, dass wir Bilder gar nicht als solche in unserem Gedächtnis behalten, sodass sie dort auch nicht überlappen und verallgemeinert werden können. So blieb nur noch, die Buchstaben- und Ziffernerkennung gewissermaßen zu imitieren, indem wir alle möglichen Varianten von Ziffern- und Buchstabenbildern nebeneinander im Gedächtnis behalten. Das ist es, was wir in der über 60 Jahre langen Geschichte der Erkennung durch NN sehen: Die Anzahl der Beispiele, die in den Trainingsdaten pro Buchstabe gespeichert sind, wuchs im Zuge dessen von einem Dutzend auf Tausende.

4. Parallel zur Vergrößerung des Trainingsdatensatzes wurde die Dimension des Vektorraums von einem Dutzend auf viele Millionen erweitert. Wozu wurde das gemacht? Eines der Hauptziele jeder Verbesserung ist die Verringerung der Anzahl der Fehler im Trainingsdatensatz nach dem Lernen. Die NN tun dies, indem sie eine Trennfläche im mehrdimensionalen Raum schaffen, die die Punkte, die zu verschiedenen Klassen gehören, voneinander trennt. In der Mathematik ist bekannt, dass für zwei **beliebige** Punktesätze eine Trennfunktion gefunden werden kann, wenn die Anzahl der Dimensionen groß genug ist. Mit zunehmender Dimension der Trennfläche wird die Oberfläche immer komplizierter und unterteilt den Raum in immer mehr getrennte kleine Bereiche. Die Trennfläche wird sehr variabel. Sie erlaubt es, die einzelnen Ziffern einer Klasse, die sich innerhalb einer Gruppe von Punkten befinden, die zu einer anderen Klasse gehören, korrekt zu erkennen. Gleichzeitig ist es möglich, dass sich ganz in der Nähe des Punktes einer Klasse ein Bereich befindet, der durch die Trennfunktion als zu einer anderen Klasse gehörig markiert wird. Wir werden es nicht wissen, bis auf der Eingabe ein Bild erscheint, das von einem Punkt in diesem Bereich dargestellt wird. Auf diese Weise wurden auch die gegnerischen Bilder entdeckt.

5. Um mit großen Datensätzen umgehen zu können, wird mehr Rechnerleistung benötigt, und dies wird

2 Bei Interesse können die Ausführungen dazu im englischen Original nachgelesen werden.

manchmal zur wichtigsten Forderung nach Verbesserungen von NN: „Alles, was wir brauchen, um dieses bisher beste Ergebnis zu erzielen, sind viele versteckte Schichten, viele Neuronen pro Schicht, zahlreiche deformierte Trainingsbilder und Grafikkarten, die das Lernen erheblich beschleunigen“ (Ciresan 2012). Damit wird die Diskussion über die Intelligenz von NNs zu jenem Zeitpunkt (etwa 2012) kommentiert, als die NNs begannen, Sprachprogramme wie ChatGPT zu entwickeln.

Fotocredit: © unsplash - Alina Grubnyak

Moderner Trend

Seit über zehn Jahren sehen wir eine neuerliche Explosion des Interesses an NN und ihrer Verwendung in vielen Bereichen. Auf der Grundlage der zeitgenössischen NN wurden die Großen Sprachmodelle (Large Language Models LLM) für die Textverarbeitung entwickelt. Der Stand der Technik der vorherigen Generation von NN für die Bilderkennung zu diesem Zeitpunkt kann wie folgt beschrieben werden:

Die Leistung der verschiedenen Erkennungsfunktionen (NN unterschiedlicher Architektur, Support Vector Machine SVM und der „K-nearest“-Algorithmus) ist fast gleich und liegt bei weniger als 1 % Fehlerquote.

Die beiden Motoren, die die Leistung der NNs in den 60 Jahren permanent vorantrieben, waren die *Größe des Trainingsdatensatzes* und die *Übertragung der Eingabedaten in höhere Dimensionen*.

Man kann davon ausgehen, dass ein LLM, das mit völlig neuem Material und neuen Zielen arbeitet, auch neue Werkzeuge erfordert. Um diese Annahme zu überprüfen, beschloss der Autor, die Antwort unter Verwendung von ChatGPT-3.5 auf Google zu finden – in dem Glauben, dass der Chat nicht gegen sich selbst aussagen wird. (Der Autor hatte in den Monaten zuvor ChatGPT-3.5 erfolgreich für das Sammeln verschiedener Informationen aus dem Internet verwendet.) ChatGPT-3.5 wurde also befragt: „Welche Mittel können zur Verbesserung der ChatGPT-Modelle beitragen?“ Die Antwort enthält eine Liste von Mitteln, und an erster und zweiter Stelle standen „Vergrößerung der Dimension des Suchraums“ und „Vergrößerung der Trainingsmenge“ - dieselben Mittel, die schon in den letzten 20 Jahren die Entwicklung der NN geleitet haben.

Hier sind die Pläne für die künftige Entwicklung von ChatGPT: „Durch die Verfeinerung der Trainingsdaten, die Förderung von iterativem Feedback und die Implementierung robuster Mechanismen zur Faktenüberprüfung können wir dazu beitragen, dass KI-Modelle wie ChatGPT weiterhin Fortschritte machen“ (De Simone 2023). Verdeutlichen wir die genannten Bedingungen für das weitere Vorankommen: Die erste ist die „Verbesserung der Trainingsdaten“, die zweite die „Verbesserung des Gradientenabstiegs“. Beide setzen die Hauptentwicklungslinie fort, die seit den Anfängen der NN-Geschichte verfolgt wird. Die dritte Bedingung ist die „Faktenüberprüfung“, d. h. die Nachbearbeitung.

Die oben aufgelisteten Eigenheiten der NN für die Bilderkennung bleiben also die gleichen, wenn NN für die Textverarbeitung verwendet werden. Dann ist es allerdings auch logisch anzunehmen, dass die Krankheiten der ChatGPT die gleichen sind wie bei den vorherigen NN.

Wir wiederholen also die Fehler, die bereits in den vorherigen NN entdeckt wurden. Es sind zwei: keine Verallgemeinerung und das Vorhandensein von gegnerischen Objekten.

Lassen Sie uns zunächst die Definitionen der Begriffe erörtern, um die es in diesem Beitrag geht. Der allgemeinste Begriff, den wir diskutieren, ist *Intelligenz*. Alle Autoren, die diesen Begriff diskutieren, führen mehrere notwendige, aber nicht hinreichend inhärente Merkmale

des Begriffs an. Ohne die Definition der Intelligenz ist es unmöglich zu beweisen, dass NN intelligent sind. Wir können allerdings beweisen, dass NN nichts besitzen, was dem Begriff der Intelligenz entspricht. Dafür reicht es aus, zu beweisen, dass NN auch nur eines der für Intelligenz notwendigen Merkmale nicht besitzen (z. B. die Fähigkeit zur Verallgemeinerung oder Verständnis).

Im *APA-Wörterbuch der Psychologie* heißt es dazu (übersetzt): „Verallgemeinerung ist der Prozess der Ableitung eines Konzepts, eines Urteils, eines Grundsatzes oder einer Theorie **aus einer begrenzten Anzahl spezifischer Fälle** und deren Anwendung in einem größeren Rahmen, oft auf eine ganze Klasse von Objekten, Ereignissen oder Menschen“ (APA Dictionary). Das bedeutet, dass ChatGPT, das auf dem Streben aufgebaut ist, eine unbegrenzte Anzahl von Fällen zu lernen, nicht in der Lage war und sein wird, zu verallgemeinern, und daher **niemals Intelligenz besitzen wird**.

Fotocredit: © unsplash - Teena Lalawat

Da Texte eine solche inhärente Eigenschaft wie *Sinn* haben, ist dessen Erkennen eine weitere notwendige Bedingung für Intelligenz. Das Wort Intelligenz kann nicht kurz beschrieben werden. Werfen wir einen Blick auf die Genealogie des Wortes:

Das Wort *Intelligenz* leitet sich von den lateinischen Substantiven *intelligentia* und *intellectus* ab, die wiederum von dem Verb *intelligere*, begreifen, abstammen. Begreifen wird dabei oft als tiefergehend gesehen als verstehen. Verstehen - **eine Fähigkeit zu begreifen**.

Das bedeutet, dass die Behauptung, dass „ein von einem intelligenten Akteur erstellter Text nicht sinnlos sein kann“, durchaus vernünftig ist. Hier ist ein Beispiel:

Die Frage an ChatGPT-3.5 lautete: **Was ist das Ergebnis eines Zusammenstoßes zwischen einem Stuhl und einer Religion?**

Die erste ChatGPT-3.5-Antwort war:

> Metaphorisch interpretiert könnte die Kollision zwischen einem Stuhl und einer Religion einen Zusammenprall zwischen weltlichen und religiösen Überzeugungen oder Werten symbolisieren. Dieser Zusammenstoß könnte sich als Konflikt zwischen verschiedenen Weltanschauungen, Überzeugungen oder Praktiken manifestieren.
> Auf einer buchstäblichen Ebene würde der Zusammenstoß zwischen einem Stuhl und einer Religion wahrscheinlich zu einer physischen Beschädigung des Stuhls und möglicherweise zu symbolischen oder metaphorischen Implikationen in Bezug auf den spezifischen religiösen Kontext führen. Bei einer religiösen Zeremonie oder einem religiösen Ritual könnte der Zusammenstoß beispielsweise symbolisch im Kontext des Glaubens und der Praktiken der jeweiligen Religion interpretiert werden. In jedem Fall wäre das Ergebnis in hohem Maße subjektiv und abhängig von individuellen Perspektiven, Überzeugungen und Interpretationen.

Es ist offensichtlich, dass die Frage keinen Sinn hat, aber der Chat hat dies nicht erkannt, was für einen intelligenten Akteur unmöglich ist. Alle fünf Sätze der Antwort wurden vom Chat generiert, und keiner dieser fünf Sätze hat einen Sinn. Das zeigt, dass der Chat nicht zwischen sinnvollen und sinnlosen Aussagen unterscheiden kann, weder im Frage- noch in seinem eigenen Antwort-Text.

ChatGPT-3.5 wurde die gleiche Frage ein zweites Mal gestellt. Nun lautete die Antwort:

> Wenn wir eine buchstäbliche Kollision zwischen einem physischen Stuhl und dem Konzept der Religion in Betracht ziehen, wäre das Ergebnis wahrscheinlich eine Beschädigung des Stuhls und je nach Kontext möglicherweise eine symbolische oder metaphorische Interpretation.
> Zum Beispiel:
> Je nach der Stärke des Aufpralls kann der Stuhl zerbrechen oder beschädigt werden.
> Symbolisch könnte der Zusammenstoß einen Zusammenstoß zwischen säkularen und religiösen Ideologien oder den Einfluss des Säkularismus auf religiöse Einrichtungen darstellen.

> In einem metaphorischen Sinne könnte der Zusammenstoß eine Infragestellung traditioneller religiöser Überzeugungen oder Praktiken oder eine Infragestellung der religiösen Autorität darstellen.
> Es ist jedoch wichtig zu beachten, dass dieses Szenario rein hypothetisch und symbolisch ist. **In Wirklichkeit kann das Konzept der Religion nicht physisch mit einem Stuhl kollidieren.**

Der zweite Text hat nicht mehr Sinn als der erste, außer dem letzten Satz: **In Wirklichkeit kann das Konzept der Religion nicht physisch mit einem Stuhl kollidieren.** Dies wäre die 100%ig intelligente Antwort auf die gestellte Frage. Dies vervollständigt unser Bild vom Intelligenzgrad von NN: NN können weder die sinnlosen Aussagen erkennen noch die sinnvollen.

Es ist erwähnenswert, dass die vom NN gefundene korrekte Aussage aus zwei Blöcken besteht, die in den Texten im Internet vorkommen: **„in Wirklichkeit das Konzept der Religion“** und **„kann nicht physisch mit kollidieren“**. Das ist für die großen Sprachmodelle (LLM) so zu erwarten. Daraus kann man schließen, dass LLM nicht verallgemeinern können.

Der halluzinierende Chatbot

Der zweite Fehler, den wir schon in den NN zur Bilderkennung finden, ist das Vorhandensein von gegnerischen Bildern (siehe oben), die man als „täuschende“ Bilder bezeichnen könnte. Da im LLM die Verarbeitung der Daten die gleiche ist wie bei den NN für die Bilderkennung, nehmen wir an, dass wir den gleichen Fehler im LLM sehen.

Nutzer, die LLM verwenden und untersuchen, haben ein Phänomen namens „Halluzination“ entdeckt, womit man eine von einem NN erzeugte Antwort mit falschen oder irreführenden Informationen bezeichnet. Zum Beispiel (aus Wikipedia) könnte ein halluzinierender Chatbot, wenn er gebeten wird, einen Finanzbericht für ein Unternehmen zu erstellen, fälschlicherweise angeben, dass der Umsatz des Unternehmens 13,6 Milliarden Dollar beträgt (oder eine andere Zahl, die scheinbar „aus der Luft gegriffen“ ist). Heute ist die allgemeine Meinung, dass „obwohl die Forschungsgemeinschaft große Anstrengungen unternommen hat, empirische Methoden zur Messung und Abschwächung von Halluzinationen zu entwickeln, unser Verständnis von LLM-Halluzinationen begrenzt bleibt.“. Gleichzeitig wird in der detaillierten Überprüfung der Leistung von LLM festgestellt, dass – „da LLM sehr flüssige und überzeugende Antworten erzeugen – ihre Halluzinationen schwieriger zu identifizieren sind und eher schädliche Folgen haben“ (Ziwei 2024).

Unsere Analyse zeigt, dass die Ursache für dieses gefährliche Phänomen eine Kette von Fehlentscheidungen ist, die bereits zu Beginn der Entwicklung getroffen wurden. Die falsche Wahl der Eingabeparameter und die Wahl des Maßes der Nähe. Um die Anzahl der Erkennungsfehler zu verringern, wurde beschlossen, die Dimension des Raums zu erhöhen. Aber der Preis dafür war hoch: die Trennfläche, die die Klassen trennt, wurde reich an hohen Frequenzen, der Bereich um jeden Punkt im Trainingssatz ist sehr klein und es besteht die Möglichkeit, dass in der Nähe eines Punktes einer Klasse ein Bereich erscheint, der zu einer anderen Klasse gehört.

Wenn man bedenkt, dass der Nutzer *getäuscht wird*, wenn der Chatbot eine *falsche oder irreführende* Nachricht (die Halluzination) produziert („gegnerischer Angriff“), kann man zu dem Schluss kommen, dass beide NN-Generationen unter demselben Fehler leiden. Die NN-Gemeinschaft versucht die Verwendung des Begriffs „gegnerisch“ zu vermeiden und ersetzt ihn durch den menschlicheren Begriff *„Halluzination“,* um die Ähnlichkeit der NN mit dem menschlichen Gehirn zu betonen.

Turing Test, Verallgemeinerung, Intelligenz

Die Entwicklung von KI wurde immer von Intelligenztests begleitet, die die Ähnlichkeit mit dem menschlichen Geist messen sollten. Meistens griff man zu diesem Zweck auf den Turing-Test zurück. In letzter Zeit jedoch, nachdem die Großen Sprachmodelle (LLM) aufgetaucht sind, wurde der Turing-Test als nicht mehr angemessen befunden. Der letzte Bericht von Ma und Mundel (Ma 2023) gibt einen guten Überblick über die Aktivitäten in diesem Bereich. Die Schlussfolgerung war: „Wir müssen uns daran erinnern, dass diese Modelle nicht wie Menschen denken. Die Fähigkeiten, die diese Modelle zeigen, erreichen sie über ihre eigenen (black-boxed) Wege“. Daher empfiehlt der Bericht, „die Fähigkeiten, die LLM haben und Menschen nicht haben, nicht nur anzuerkennen, sondern auch zu untersuchen“.

Vielleicht ist es ein guter Ratschlag, die Black-Box in Ruhe zu lassen und sich darauf zu konzentrieren, den praktischen Nutzen von LLM zu erweitern, aber der Ansturm auf die intellektuelle Superkraft des LLM nimmt zu. Allerdings kann es zu seltenen, aber sehr teuren Fehlern kommen. Es ist nicht wichtig, wie viel man weiß. Die Frage ist, wie

man die Fehler handhabt. Eine der Fragen, die jemanden als intelligent disqualifizieren, ist, ob er in seinem Wissensgebiet sinnlose Aussagen produziert und solche nicht erkennt. Wie etwa in der Mathematik, wo ein einziges Beispiel, das der Theorie widerspricht, genügt, um die Theorie zu verwerfen. Die modernen NN erfüllen diese beiden Kriterien der Intelligenz nicht: die Fähigkeit, eine vernünftige Aussage von einer unvernünftigen zu unterscheiden, und die Fähigkeit zur Verallgemeinerung.

Es gibt einen merkwürdigen Mythos im Land der KI – es sei unmöglich zu verstehen, wie die NN funktionieren. Es gibt zwei Gründe für das Auftauchen und die Verbreitung dieses Mythos: 1) Oben haben wir erklärt, wie die NN funktionieren und warum verschiedene Arten von Fehlern auftreten: es zeigt die Tatsache, dass NN nicht das Denken simulieren, 2) Die Mitglieder der KI-Gemeinschaft unterstützen diesen Mythos, weil sie ihn benutzen, um die Ähnlichkeit zwischen den NN und dem menschlichen Gehirn zu behaupten, nach dem Motto: „Das Gehirn ist sehr komplex und wir wissen sehr wenig darüber, wie es funktioniert"; „die NN sind auch sehr komplex und deshalb wissen wir auch nicht, wie sie funktionieren." Diese Undurchschaubarkeit wird zu einem zusätzlichen Argument für die Behauptung der Ähnlichkeit zwischen dem Gehirn und den NN.

Abschließende Bemerkungen

1. Die von LLM erstellten Texte lesen sich sehr menschlich, aber das reicht nicht aus, um sie als vernünftig zu qualifizieren. Es ist bekannt, dass Studenten manchmal nicht lernen, sondern pauken – sie kennen die richtigen Antworten auf eine begrenzte Anzahl von Fragen und erhalten gute Noten in der Prüfung, aber sie verstehen die Bedeutung ihrer Antworten nicht.

2. Vor sechzig Jahren veröffentlichte M. Bongard das Buch „Erkennungsprobleme" (Bongard 1970), an dessen Ende 100 grafische Rätsel zum Thema Verallgemeinerung stehen. Diese Aufgaben waren für den Menschen nicht einfach. Die Frage war: Können wir ein Programm entwickeln, das die richtigen Antworten auf all diese Rätsel findet? Damals war man sich einig, dass ein solches Programm über Intelligenz verfügen würde (ohne diesen Begriff zu definieren). Fünf Jahre später veröffentlichte einer von Bongards Mitarbeitern, V. Maximov, einen Algorithmus, der das Problem löste, aber niemand schrie „Halleluja!" oder „Eureka" – niemand dachte, dass „wir es geschafft haben". Dann kam die Einsicht, dass (zumindest für heute) diese Intelligenz ein Wunder sein muss. Die großartige russische Dichterin Anna Achmatova schrieb (übers.):

> Wenn du wüsstest, was für ein Kauderwelsch der Vers ist,
> der wächst, von allen Schändlichkeiten befreit,
> wie die gelben Sommerblumen des Löwenzahns,
> wie die Klette und wie das Quinoa-Kraut.

Sie sagte: Wenn man erst einmal weiß, wie schlicht die Dinge und Gedanken sind, die die hohe Poesie hervorbringen, wird man vielleicht nicht mehr so begeistert von der Poesie sein.

Ähnlich scheint es sich mit Texten zu verhalten: Wenn wir verstehen, wie menschenähnliche Texte entstehen können, ist die Aura des Wunders verschwunden, und wir sind bereit, den aufrührerischen Gedanken zu akzeptieren, dass dieser recht vernünftige Text nach einigen klaren und einfachen Regeln und ohne Einsatz des Verstandes entstanden ist. Nachdem wir diesen kühnen Schritt getan haben, müssen wir den nächsten wagen und die Frage stellen: „Sind alle Texte, die von Menschen geschrieben oder gesprochen werden, Produkte des „produktiven Denkens" (im Sinne von Max Wertheimer), d. h. durch das Sammeln und Verarbeiten von Informationen, **oder** sind einige nach bestimmten Rezepten entstanden. Die schnelle und weite Verbreitung von Chats deutet nicht auf ein hohes intellektuelles Niveau der LLM hin, sondern eher auf niedrige Anforderungen an Texte.

Die Diskussion über die Frage „Denken Menschen?" ist nicht unsinnig - sie hat eine lange Geschichte.

Die berühmteste Episode in dieser Diskussion fand 1925 statt, als Bertrand Russell feststellte, dass **„die meisten Menschen eher sterben würden, als zu denken – und das tun sie auch".** Es muss betont werden, dass Russell ein hoch angesehener Philosoph und Wissenschaftler war, und er tat dies nicht auf einer Pressekonferenz, sondern in seinem Buch „Das ABC der Relativitätstheorie", in dem er Einsteins Theorie erklärte. Im nächsten Monat druckte die populäre Londoner Zeitung „The Observer" Russells Bemerkung in einer Auswahl mit dem Titel „Sprüche der Woche" ab. Tatsächlich erschien und kursierte dieser Spruch im 20. Jahrhundert und ist auch im 21. Jahrhundert nicht vergessen: 2021 wurde ein Buch mit Russells Spruch als Titel veröffentlicht.

Es handelt sich hier nicht um eine abstrakte Diskussion. In einem Be-

richt über das Bildungswesen in den USA (D. Halpern „Productive thinking in psychology“) heißt es: „In den letzten 15 Jahren hat die Fähigkeit amerikanischer Schüler, zu denken (und sich nicht nur zu erinnern), deutlich abgenommen“. Das Denken wurde durch Pauken ersetzt - genau wie in den NN. Eine der drei Harvard-Studienstrategien für Studenten im ersten Studienjahr lautete einmal: „Pauken am Abend vor einer Prüfung ist keine effektive Strategie“. Heute wird das tragische Ereignis der Ersetzung des Denkens durch Auswendiglernen von der KI-Gemeinschaft verkündet und von den Massenmedien als großer Schritt zur Intelligenz dargestellt.

Schlussfolgerung

Weder die NNs für die Bilderkennung noch die LLM haben die Funktionalität, um als intelligent gelten zu können.

Danksagungen

Dieser Artikel wäre nie geschrieben worden, wenn ich nicht seit 30 Jahren zahlreiche Gespräche mit Alexander Pashintsev über diese Themen geführt hätte.

Literatur

American Psychological Association Dictionary. https://dictionary.apa.org/generalization

Bongard, M.M. (1970): Pattern Recognition. New York: Spartan Books.

Baldominos A. et.al. A Survey of Handwritten Character Recognition with MNIST and EMNIST. Appl. Sci. 2019, 9, 3169; doi:10.3390/app9153169.

Ciresan D. , Schmidhuber J. Multi-column Deep Neural Networks for Image Classification. 978-1-4673-1228-8/12/$31.00 ©2012 IEEE.

De Simone V. (2023). Ein Überblick über den Einsatz von KI/ML in der Fertigung. Proced Computer Science, Band 217 2023, Seiten 1820-1829 (https://www.sciencedirect.com/journal/procedia-computer-science/vol/217/suppl/C)

Guberman, S. (2007): Unorthodox geology and geophysics. Polimetrica.

Le Cun, Yann et al (1989): Handwritten Digit Recognition with a Back-Propagation Network. In: Advances in neural information processing systems 2NIPS 1989, 396-404. https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf

LeCun Yann; Bottou L.; Bengio Y. & Patrick Haffner (1998): Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Ma M., Mandal J. Overcoming Turing: Rethinking Evaluation in the Era of Large Language Models. Stanfors CodeX, 2023. https://law.stanford.edu/2023/11/16/overcoming-turing-rethinking-evaluation-in-the-era-of-large-language-models/

McCulloch, Warren S. & Walter H. Pitts (1943): A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133. http://dx.doi.org/10.1007/BF02478259.

OpenAI(2023): GPT-4 Technical Report. Xiv:2303.08774v4 [cs.CL] 19 Dec 2023

Rosenblatt, Frank (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408.

Turing, Alan M. (1950): Computing Machinery and Intelligence. Mind, 49, 433-460