vordenker
April 2024

Künstliche Dummheit – Nicht-triviale Maschinen und der Fluch der Rekursion

Joachim Paul

How to cite:
Joachim Paul; Künstliche Dummheit – Nicht-triviale Maschinen und der Fluch der Rekursion
online: www.vordenker.de Neuss 2024, J. Paul (Ed.), ISSN 1619-9324
URL, pdf: < https://www.vordenker.de/jpaul/KI_NTM_Varianz_Rekursion.pdf >
alternativ via Wordpress,
URL: <https://www.vordenker.de/blog/2497/kuenstliche-dummheit-nicht-triviale-maschinen-und-der-fluch-der-rekursion/ >

vordenker
ISSN 1619-9324

Künstliche Dummheit
– Nicht-triviale Maschinen und der Fluch der Rekursion

Joachim Paul

Seit Mai 2023 mehren sich nun die Indizien, dass neben dem Marketing Buzzword „Künstliche Intelligenz" in Zukunft wohl auch der künstlichen Dummheit ein Platz in den Berichterstattungen eingeräumt werden muss. Auf Heise online war gar von Demenz die Rede. Das ist zwar ebenso wie „künstliche Dummheit" und „künstliche Intelligenz" ein ausgeprägter und kritikwürdiger Anthropomorphismus, möglicherweise treffend daran ist aber die Tatsache, dass Demenz etwas mit Alterungsprozessen zu tun hat. Diese bilden sich bei IT-Systemen auch in der Versionierung ab.

Das erste Auftauchen des o.g. Begriffs „Künstliche Intelligenz, KI, oder AI, und die damit verbundene Motivation lässt sich präzise angeben. John McCarthy garnierte 1956 sein Dartmouth Summer Research Project on Artificial Intelligence, kurz Dartmouth Conference, mit dem neuen Begriff AI, um Forschungs- und Spesengelder einzuwerben. Man interessierte sich auf dieser Konferenz vorzugsweise für symbolische Verfahren und grenzte sich dadurch von Norbert Wiener und seinen Kybernetik-Kollegen der Macy-Konferenzen ab, die zumeist konnektionistische, auf Verbindungen von Knoten basierende Modelle wie Perzeptronen und neuronale Netzwerke in den Blick genommen hatten.

In den 90ern nahm das Thema KI erneut Fahrt auf, nach dem sogenannten KI-Winter hatten dieses Mal die heuristischen Modelle, die ANNs (artificial neural networks) die Nase vorn. Deep Learning Networks, d.h. ANN mit mehreren internen Layern zwischen Eingabe- und Ausgabeschicht, erwiesen entgültig ihre Überlegenheit gegenüber den symolischen Ansätzen, nachdem der riesige verschlagwortete Bilddatensatz ImageNet durch die Informatikerin Fei-Fei Li und ihr Team 2009 bereitgestellt wurde. Der unbestreitbar größte Durchbruch in den Augen der Weltöffentlichkeit erfolgte dann am 30.11.2022, als OpenAI sein Large Language Model (LLM) Generative Pretrained Transformer, GPT Version 3 als Dialogsystem ChatGPT zur öffentlichen Nutzung freischaltete.

Denn im Mai 2023 veröffentlichte eine internationale Gruppe aus sechs Wissenschaftlern von fünf Universitäten, darunter Oxford, Cambridge und Toronto, eine Studie mit dem Titel „The Curse of Recursion: Training on Generated Data Makes Models Forget", in dem für den Fall des Trainings mit von durch KIs produzierten „künstlichen" Datensätzen die Existenz degenerativer Prozesse während der Trainingsphase in einer ganzen Reihe von Modelltypen nachgewiesen und demonstriert wird. Neben den LLMs sind dies GMMs (Gaussian Mixture Models zum Sortieren, bzw. Clustern von Daten) und VAEs (Variational Autoencoders zur Erkennung von handgeschriebenen Ziffern). Die Autoren sprechen vom Modellkollaps und liefern auch gleich ein Rezept zur Vermeidung des Zusammenbruchs. Es müsse sichergestellt werden,

das von Version zu Version einer KI ausschließlich von Menschen produziertes Trainingsmaterial verwendet wird.

Ende Oktober 2023 erscheint eine Arbeit von drei Forschern der Universitäten Stanford und Berkeley, die sich ausschließlich auf den Textgenerator ChatGPT und seine Versionen 3.5 und 4 bezieht, „How Is ChatGPT's Behavior Changing over Time?" Beiden Versionen des populären Systems wurden insgesamt sieben verschiedene Arten von Aufgaben zu zwei verschiedenen Zeitpunkten, März und Juni 2023, gestellt.

Die Performances fielen dabei abhängig von den Zeitpunkten signifikant unterschiedlich aus. So identifizierte GPT4 im März Primzahlen und zusammengesetzte Zahlen zu 84% richtig, im Juni waren dies nur noch 51%. GPT3.5 hingegen war in dieser Aufgabe im Juni deutlich besser als im März.

Ein unvoreingenommener Beobachter mag sich zwar die Frage stellen, ob die Identifizierung von Primzahlen ein geeigneter Job für Textgeneratoren ist, gleichwohl ist das nur ein Beispiel. Denn auch für die anderen Aufgaben-stellungen bestätigen die Untersuchungen die Existenz von Unklarheiten bezüglich der Auswirkungen von Updates und von außen durchgeführten Variationen einiger Netzparameter.

Die Autoren weisen explizit darauf hin, dass das Wie und Wann der Updates und Aktualisierungen von LLMs wie ChatGPT nicht transparent ist und kündigen ein Langzeitprojekt an. Sie kommen zunächst zu dem Ergebnis, dass die Verbesserung der Leistung des Modells bei einigen Aufgaben, z.B. durch auf bestimmte zusätzlich zur Feinabstimmung herangezogene Daten unerwartete Nebeneffekte auf das Modellverhalten bei anderen Aufgabenstellungen haben kann.

An dieser Stelle sei eine spekulative Frage erlaubt. Sind die sich nach einer Feinabstimmung für bestimmte Aufgaben sich ergebenden Verschlechterungen der Performance anderer Aufgabenstellungen möglicherweise ein Indiz dafür, dass das neuronale Netz zu klein ist, bzw. rein quantitativ zu wenig Parameter enthält, um beide Aufgabenfelder erfolgreich bearbeiten zu können? Denn wenn das so ist, dann sind diese Feinabstimmungen wenig mehr als bloßes Tricksen, bzw. Herumprobieren.

Entsprechend dem Hype-Thema KI folgt beiden Veröffentlichungen ein Rauschen im Blätterwald der Feuilletons, der Fachpresse und der einschlägigen Blogs. Es sei sehr wahrscheinlich, heißt es, dass KIs in Zukunft immer häufiger auch mit Daten trainiert werden, die selbst Outputs von KI-Systemen sind.

Bald macht das Wort von der KI, die ihren eigenen Schwanz frisst, die Runde. Im Popular Mechanics Magazine fordert der Autor Darren Orf seine Leser erstmal auf, zu Popcorn zu greifen und bemüht den Ouroboros, die sich selbst in den Schwanz beißende Schlange der Ewigkeit, als sprachliches Bild.

Schon zuvor im August spricht Gary Marcus, ein populärer Kritiker des KI-Hype, gar von der durch LLMs getriebenen „enshittification" des Internet.

„Garbage in, garbage out. Data pollution is ruining generative AI's future", kommentiert Ben Lutkevich auf TechTarget.

Datenverschmutzung? Müll rein, Müll raus? Aber der Output ist zu Beginn in der Regel kein Müll, bzw. wird nicht als solcher aufgefasst. Daher ist diese Erklärung allein möglicherweise zu einfach. Das Training von neuronalen Netzen mit Outputdaten von neuronalen Netzen hat erstens eine Abnahme der Varianz der im Netz gespeicherten Wichtungen zur Folge und stellt im Prinzip eine Analogie dar zur Qualitätsabnahme bei Fotokopien von Fotokopien. Darüber hinaus schlägt hier ein theoretisch ausformuliertes Prinzip zu, dass schon seit 1962 bekannt ist und für alle Arten von Maschinen mit endlich vielen internen Zuständen gilt.

Eine sogenannte „finite state machine" ist ein abstraktes mathematisches Konzept, das durch eine Black Box mit einem Input und einem Output , z.b. für alphanumerische Zeichen, beschreibbar ist und deren Prinzip an folgenden beiden Beispielen erläutert werden kann.

Der aktuelle Output einer solchen Maschine ist eine einer feststehenden Regel folgende Funktion des Inputs. Jedes Zeichen als Input hat ein entsprechendes Zeichen als Output zur Folge und es lässt sich eine Liste der Input-Output-Zeichenpaare aufstellen. Diese Art der Maschine wird trivial genannt. Ein Beispiel ist z.B. die Lenkung eines PKW. Input: Lenkrad nach links, Output: Fahrzeug fährt nach links, geradeaus/ geradeaus, rechts/ rechts, etc.

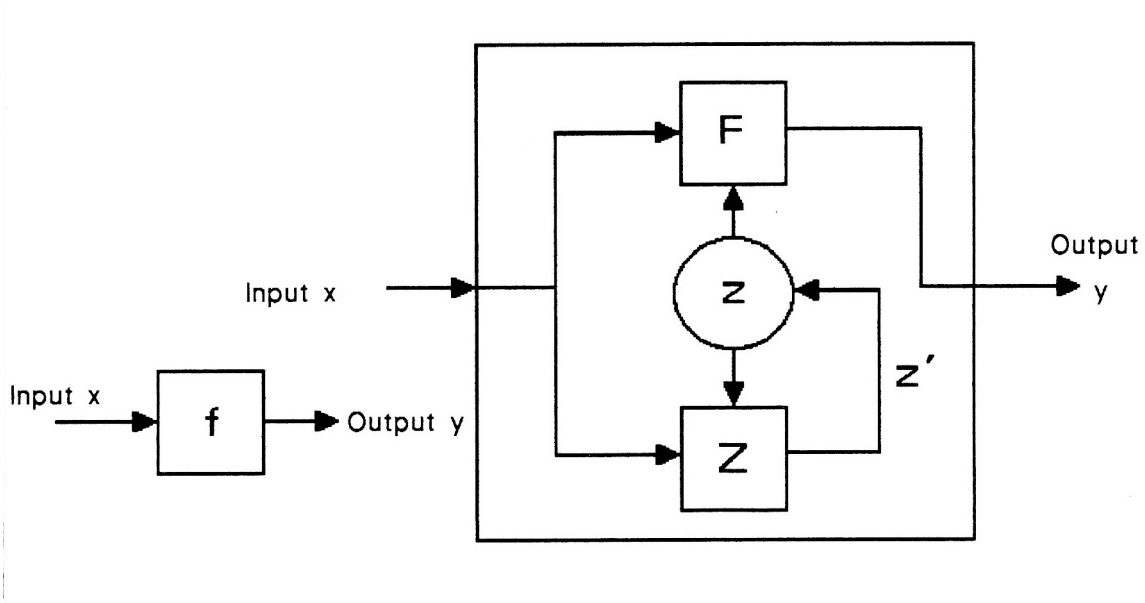

Schaubild 1: Links: Triviale Maschine, y = f(x) // Rechts: Nichttriviale Maschine mit interner Zustandsfunktion Z: y = F(x,z) mit z' = Z(x,z) (frei nach H. v. Foerster)

Eine andere Art einer „finite state machine" ergibt sich, wenn sie einen zusätzlichen internen Zustand besitzt, der ebenfalls einer Regel folgend abhängig vom Input ist. Der Output ist nun eine dem Beobachter unbekannte aber einer festen Regel folgende Funktion des aktuellen Inputs und des aktuellen internen Zustands. Jeder Input ändert aber ebenfalls einer festen Regel folgend den internen Zustand der Maschine. Wird nun in einem nächsten Schritt ein weiterer Input angelegt, ist der Output ein Ergebnis des jetzt aktuellen Inputs und des internen Zustands, der ja im vorangegangenen Schritt verändert wurde. Das nun vorliegende Konstrukt nennt man nicht-triviale Maschine, kurz NTM. Der Output einer sochen NTM ist für einen gegebenen Input nicht mehr vorhersagbar, d.h., die NTM kann analytisch nicht bestimmt werden, sie ist nicht determinierbar. In Bezug auf ihre Konstruktion ist sie ist lediglich synthetisch determiniert. Da der interne Zustand einer NTM sich mit jedem neuen Input ändert und dies wiederum Einfluss auf den Output des nächsten Schrittes hat, wird die NTM geschichtsabhängig genannt. Kompliziertere Algorithmen sind grundsätzlich nicht-triviale Maschinen.

Es leuchtet unmittelbar ein, dass auch deep learning neural networks in diese Kategorie von Maschinen gehören. Zwar handelt es sich hier um extrem viele mögliche interne Zustände, GPT3 hat 175 Mrd. veränderliche Parameter und bei GPT4 sind dies über eine Billion. Gleichwohl sind das immer noch endlich viele Zustände, denn auch die rechnerische Darstellung von Fließkomma-zahlen, hier die synaptischen Wichtungen im ANN, ist ganz prinzipiell auf eine endliche Anzahl von Binärstellen beschränkt.
Wenn man nun mehrere solche NTM via Netzwerk zusammenschaltet - entsprechend kann man sich auch KI-Systeme und ihre Inputs, Outputs und Trainingsdatenpools als via Internet zusammengeschaltet denken -, dann kommt dabei ein neues System heraus, das nun wiederum als eine einzige NTM betrachtet werden kann, deren Verhalten ebensowenig vorhergesagt werden kann.

Urheber der Theorie der „finite state machines" ist der amerikanische Elektroingenieur Arthur Gill (1930-2020) aus Berkeley, der 1962 seine Einführung dazu veröffentlichte.[1] Der Kybernetiker Heinz von Foerster (1911-2002) stellte Gills Unbestimmbarkeitsprinzip für nicht-triviale Maschinen an eine Seite mit Gödels Unvollständigkeitssatz und Heisenbergs Unschärferelation. Er unterstellt – das ist gleichwohl kontrovers diskutierbar - Gills Konzept eine größere Allgemeingültigkeit als dem Konzept der Turing-Maschine.[2]

Gills NTM zeigen bei Rückkopplung ihres Outputs auf ihren Input, also bei einer Situation vergleichbar der des Trainings einer KI mit KI-produzierten Daten, ein grundsätzliches sowie seltsames Verhalten. Von Foerster demonstrierte dies um 1970 am Beispiel einer ganz simplen nicht-trivialen Maschine mit nur zwei möglichen internen Zuständen. Koppelt man hier den Output rekursiv auf den Input zurück, dann produziert eine solche NTM bereits nach wenigen Schritten eine sich ewig wiederholende Output-Folge, die heute „seltsamer Attraktor"

genannt wird.[3] Die Output-Werte der Folge sind die Eigenwerte dieser speziellen nicht-trivialen Maschine.

Und aufgrund der vorliegenden Studien besteht die berechtigte grundsätzliche Annahme, dass seltsame Attraktoren zum natürlichen Verhalten künstlicher Intelligenzen gehören.

Aber womöglich kann KI mit Hilfe von KI gerettet werden. Der Student Tom Tlok von der Hochschule Wedel entwickelte im Rahmen seiner Masterarbeit ein Werkzeug, dass Texte zu erkennen vermag, die mit Hilfe von künstlicher Intelligenz geschrieben wurden, Trefferquote ca. 98 Prozent.

Ein Backpropagation-Algorithmus bewertet Outputs von Backpropagation-Algorithmen. Auch eine Form von Rekursion.

Joachim Paul, Neuss im April 2024

Quellen:

[1] Arthur Gill, Introduction to the Theory of Finite State Machines, New York: McGraw-Hill 1962

[2] Heinz von Foerster, Principles of Self-Organization - In a Socio-Managerial Context,
in: Self-Organization and Management of Social Systems, eds. H. Ulrich, G.J. B. Probst, Berlin Heidelberg 1984, p. 2-24

[3] Heinz von Foerster, Molecular Ethology. An Immodest Proposal for Semantic Clarification, in: Molecular Mechanisms in Memory and Learning, ed. G. Ungar, New York 1970, p. 213-248.
dt.: Molekular-Ethologie: ein unbescheidener Versuch semantischer Klärung,
in: Heinz von Foerster, Sicht und Einsicht, Versuche zu einer operativen Erkenntnistheorie, Wiesbaden 1985, S. 173-204

Links im Text:
https://www.heise.de/news/Kuenftige-KI-Modelle-potenziell-von-Demenz-bedroht-9209900.html

https://arxiv.org/abs/2305.17493

https://arxiv.org/abs/2307.09009

https://www.popularmechanics.com/technology/a44675279/ai-content-model-collapse/

https://garymarcus.substack.com/p/the-imminent-enshittification-of

https://www.techtarget.com/whatis/feature/Model-collapse-explained-How-synthetic-training-data-breaks-AI

https://en.wikipedia.org/wiki/Finite-state_machine

https://www.ndr.de/nachrichten/schleswig-holstein/ChatGPT-Student-aus-Wedel-entlarvt-kuenstliche-Intelligenz,kidetektor108.html