KI-Revolution frisst Unmengen Strom – bremst Deutschland seine Position bald aus?

Welt, 18.06.2023, Benedikt Fuest

https://www.welt.de/wirtschaft/plus245796990/Energie-Hunger-von-KI-Wie-Deutschlands-Politik-den-eigenen-Fortschritt-gefaehrdet.html?sc_src=email_4468534&sc_lid=456275930&sc_uid=9b9AoAfTYB&sc_llid=4419&sc_cid=4468534&cid=email.crm.redaktion.newsletter.wirtschaft&sc_eh=94c824e22aa172ca1

Das Training künstlicher Intelligenz ist extrem energieaufwendig – innovative Technologien sollen den ökologischen Fußabdruck von ChatGPT und Co. zunehmend verringern. Eine neue Regulierung allerdings könnte den Bau weiterer KI-Rechenzentren in Deutschland blockieren.

Auf den ersten Blick sieht das Innere von Europas leistungsstärkstem kommerziellem Rechenzentrum für künstliche Intelligenz (KI) unscheinbar aus: Ein halbes Dutzend mattschwarz lackierte Schränke stehen in einem Raum. Sie lassen noch viel Platz für weiteren Ausbau. Darüber laufen ein einzelnes Rohr des Kühlungssystems und einige Bündel Netzwerkkabel über Stahlträger. Erst auf den zweiten Blick fallen die Stromanschlüsse an der Decke auf. Gleich 14 dicke Starkstromkabel führen von grauen Doppel-Verteilern in jeden einzelnen Schrank, zeugen vom enormen Energiehunger der verbauten Technik.

Die 64 Server in den Schränken gehören dem Heidelberger KI-Start-up Aleph Alpha, welches hier seine selbstlernenden Sprach-Algorithmen trainiert. 512 Grafikchips vom Typ Nvidia HGX A100, acht pro Server, bilden gemeinsam einen großen Supercomputer namens Alpha One. Läuft der unter Volllast, dann benötigt jeder einzelne Server bis zu 6,5 Kilowatt, insgesamt beziehen die Nvidia-Chips allein knapp zehn Megawattstunden Strom pro Tag, was in etwa dem Verbrauch von 2000 Handys im Jahr entspricht.

Hinzu kommt der Energieverbrauch für Speicher, Netzwerk, Klimatisierung und Gebäudetechnik. „Alpha One wird ausschließlich mit zertifiziertem Strom aus umweltfreundlichen erneuerbaren Energien betrieben“, kommentiert ein Unternehmenssprecher. Was nichts daran ändert, dass KI-Anwendungen viel Energie benötigen.

Die Revolution des Arbeitsalltags durch künstliche Intelligenz steht gerade erst am Anfang – aber schon jetzt wird deutlich, dass diese Revolution erheblich ressourcenintensiver wird als der Betrieb etablierter Unternehmens-IT. Das fängt bei den Prozessoren an, die fürs Training der Algorithmen nötig sind: Klassische Computer taugen nicht für das Training der künstlichen Intelligenz, nur die speziellen Grafikprozessoren, die viele Berechnungen parallel abarbeiten können, sind schnell genug. „Grafikprozessoren sind derzeit erheblich schwieriger zu bekommen als Drogen“, kommentierte Twitter-Eigentümer Elon Musk Mitte Mai seinen Versuch, für das soziale Netzwerk einen eigenen KI-Supercomputer zu bauen.

Aleph Alphas Alpha One mit 512 Chips ist noch bescheiden im Vergleich zur US-Konkurrenz: 10.000 Grafikkarten will allein Twitter für sein KI-Training rechnen

lassen, 16.000 Grafikchips sollen im neuen Supercomputer des Facebook-Konzerns Meta Platforms arbeiten. Grafikchips sind, gemessen an der Leistungsabfrage pro Chip, die stromhungrigsten Komponenten in der Chipwelt überhaupt. Sie müssen besonders aufwendig gekühlt werden, benötigen komplexe Netzwerktechnik und besonders schnelle Speicherbausteine.

Einziger Trainingsdurchlauf emittiert bis zu 120 Tonnen CO_2

Der Wettlauf um das schnellste Training für neue KI-Algorithmen ist eröffnet – und jeder, der ernsthaft mitmachen möchte, benötigt einen eigenen Supercomputer. Dafür sind jede Menge Rohstoffe, Strom und Kühlwasser notwendig: erst für den Bau, dann für den Betrieb der neuen Generation von Rechenzentren.

„Um ein großes Modell wie GPT 3.0 in einem sinnvollen Zeitrahmen trainieren zu können, müssen Sie erst einmal etwa zehn Millionen Euro in Hardware investieren“, kommentiert Wolfgang Maaß, Forschungsbereichsleiter am Deutschen Zentrum für Künstliche Intelligenz (DFKI) in Saarbrücken, „ein Trainingsdurchlauf für ein großes Modell dauert je nach Anzahl der eingesetzten Grafikchips in etwa zwei bis vier Wochen und verbraucht über 300 Megawattstunden Strom.“ Würde man den deutschen Strommix mit etwa 400 Kilogramm CO_2-Ausstoß pro Megawattstunde annehmen, sorgt also ein einziger Trainingsdurchlauf eines großen Modells für Emissionen von etwa 120 Tonnen Kohlendioxid.

Da aber ein einziger Durchlauf nicht ausreicht, gehen aktuelle Studien von weit über 1200 Megawattstunden Energieverbrauch für das Training von GPT-3 aus. Denn jedes Mal, wenn eine künstliche Intelligenz neue Inhalte lernen soll, muss sie neu trainiert werden – schrittweises Lernen wie beim Menschen ist unmöglich, warnt eine Studie kanadischer KI-Forscher um die Wissenschaftlerin Alexandra Luccioni, die den Ressourcenverbrauch von KI untersucht haben.

Damit ist das Trainieren künstlicher Intelligenz viel aufwendiger als die Arbeit mit klassischen Datenbanken, wie sie bislang in der Unternehmens-IT eingesetzt werden. Nur die speziellen Grafikchips, die ursprünglich für die Berechnung von Computerspielen entwickelt wurden, können die Matrizengleichungen hinter den Algorithmen überhaupt in einem sinnvollen Zeitrahmen berechnen – sie müssen also nicht für jede Gleichung einen neuen Rechenschritt verwenden, sondern können eine gesamte Matrix in einem Rechenschritt lösen. „Unsere Chips sind 20 Mal effizienter als gewöhnliche Serverchips“, sagt Shar Narasimhan, Datacenter-Efficency-Experte bei Nvidia.

Doch auch die Grafikprozessoren müssen wochenlang rechnen, um ein Modell zu trainieren. „Mit jedem Parameter, um den die Modelle größer werden, steigt der Leistungsbedarf“, erklärt Narasimhan. GPT 3.0, das als „ChatGPT“ Furore machte, basierte auf 175 Milliarden Parametern. Der lNachfolger GPT 4.0 soll bereits eine Billion Parameter nutzen, der Energiebedarf für einen Trainingsdurchlauf liegt etwa 20 Mal so hoch.

Im Wettlauf der großen Internet-Konzerne zeichnet sich bereits jetzt ab, dass die KI-Modelle umso besser werden, je mehr Parameter sie berücksichtigen können. Während GPT-3 noch an einfachen Textaufgaben scheiterte, vermag GPT-4 juristische Staatsexamen zu bestehen oder Businesspläne zu schreiben.

Dementsprechend groß ist aktuell die Konkurrenz von Microsoft und den GPT-Erfindern des Start-ups OpenAI, von Google, Meta und Amazon im Rennen um die leistungsfähigsten Modelle.

Grenzen setzt allein die Zahl der verfügbaren Chips. „Geld oder Ressourcenverbrauch spielen bei diesem Wettlauf um Marktanteile aktuell überhaupt keine Rolle“, kommentiert DFKI-Experte Maaß. Das könnte sich ändern, sobald erste Modelle im Masseneinsatz sind und die Konzerne bei jeder einzelnen Anfrage – davon geht Maaß aktuell aus – Energiekosten im Cent-Bereich stemmen müssen. Denn erst im Masseneinsatz zeigen sich die Gesamtkosten der künstlichen Intelligenz: „Von nun an wird der Gewinn pro Suchanfrage für immer fallen“, kommentierte Satya Nadella, Microsoft-Chef, das Dilemma, vor dem die Konzerne nun stehen. Nicht nur das Training, auch die Berechnung der Anfragen an die künstliche Intelligenz benötigt viel Rechenleistung.

Wie viel Strom genau die KI-Berechnungen verbrauchen, verraten die Konzerne nicht. Doch Googles eigene KI-Forscher haben veröffentlicht, dass künstliche Intelligenz in den vergangenen drei Jahren etwa bei Google zehn bis 15 Prozent des gesamten Stromverbrauchs des Unternehmens verursachte. Der lag im Jahr 2021 bei 18,3 Terawattstunden.

Neue Chip-Generation deutlich effizienter

Hinzu kommt der Ressourcenverbrauch für den Bau der Rechenzentren: Das US-Start-up Hugging Face veröffentlichte im November 2022 eine Studie, die die Emissionen eines großen Algorithmus über seinen gesamten Lebenszyklus – vom Bau der Hardware über das Training bis zur täglichen Nutzung – untersuchte. Sie kam zu dem Ergebnis, dass das Training nur etwa ein Drittel der Emissionen verursacht. Hinzu kommen noch einmal dieselben Verbräuche für den Einsatz der Algorithmen nach dem Training sowie den Bau und Betrieb der Hardware.

Sowohl Google als auch Facebook und Amazon bauen deswegen aktuell eigene Chips, die insbesondere die sogenannte Inferenz-Rechnung, also die Anwendung der fertig trainierten Algorithmen, besonders energieeffizient beherrschen.

Und auch Chipbauer und Nvidia-Konkurrent Qualcomm entwickelt seinen neuen „Cloud AI 100“-Chip für die möglichst effiziente Anwendung der KI. Das Training der Modelle wird vorerst aufwendig bleiben, bei der Anwendung dagegen könnte die Energiebilanz künftig ähnlich ausfallen wie bei klassischen Suchanfragen im Netz, hoffen die Entwickler.

Aleph Alphas Rechenzentrum Alpha One ist, verglichen mit den Plänen der US-Konzerne, bislang bescheiden. Dass die Planer Raum für Zuwachs gelassen haben, ist aber kein Zufall. Aktuell bemüht sich das Start-up um weitere Investorengelder, um den Wettlauf für das Training der KI mitgehen zu können.

Doch beim Ausbau ihrer Rechenzentren könnte den deutschen Betreibern künftig eine neue Umweltschutz-Gesetzgebung der Bundesregierung im Weg stehen, warnt der Eco-Verband der Internetwirtschaft: „Um künstliche Intelligenz in Deutschland wettbewerbsfähig zu halten, benötigen wir neue, leistungsfähigere Rechenzentren“, sagt Eco-Chef Alexander Rabe. „Doch das Wirtschaftsministerium unter Habeck

plant aktuell ein Energieeffizienzgesetz, dessen Vorgaben einem Neubaumoratorium für KI-Rechenzentren gleichkommen.“

Die Öko-Strategen des Wirtschaftsministeriums wollen den Betreibern der Rechenzentren vorschreiben, dass sie die Abwärme ihrer Hardware nutzen, etwa in Fernwärmenetze einspeisen. „Doch solche Abwärmesenken gibt es längst nicht überall, im Gegenteil: Die allermeisten Stadtwerke winken ab“, sagt Rabe.

Im Sommer, wenn die Server viel kühlen müssen, ist kein Bedarf, im Winter dagegen ist die Temperatur der Abwärme nicht hoch genug für Fernwärme. „Damit gefährden wir zum einen die Wettbewerbsposition der deutschen KI-Start-ups – zum anderen rauben wir uns Möglichkeiten der Regulierung dieser Technologie“, so der Eco-Chef. Man sollte sich nichts vormachen, warnt er: „Nur da, wo die Server stehen, kann man auch Regeln für die Nutzung der KI durchsetzen.“