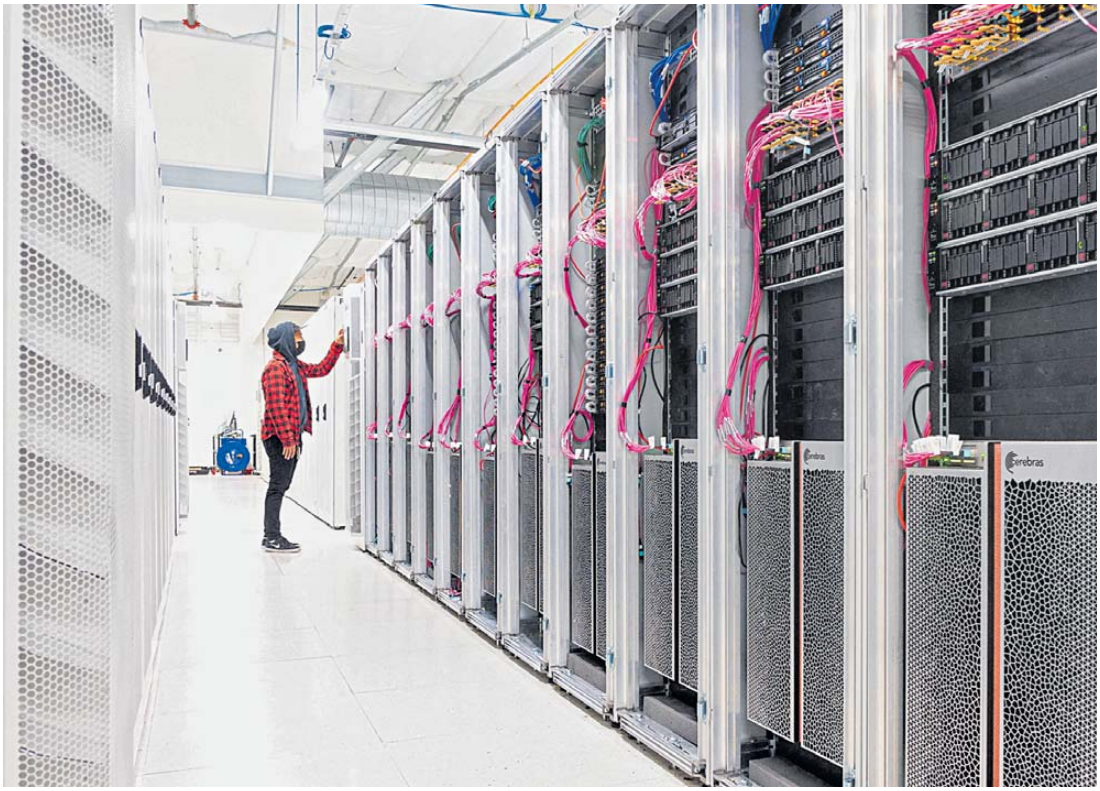

Der kalifornische Supercomputer Andromeda. Der Energieverbrauch solcher Maschinen steigt mit ihrer Leistungsfähigkeit. FOTO: REBECCA LEWINGTON/REUTERS

Der Energiehunger der KIs

Wie viel Strom Chat-GPT und andere intelligente Bots verbrauchen, wissen allein die Techfirmen. Doch neuere Berechnungen lassen vermuten, dass ihr Bedarf ständig zunimmt

Von Tobias Landwehr

Brute Force Learning ist simpel und brachial. Eine Künstliche Intelligenz (KI) lernt dabei, indem sie Millionen Inhalte in sehr vielen Wiederholungen durchsieht. Mit hinreichender Ausdauer führt das zum erwünschten Ergebnis, etwa der Fähigkeit, lesbare Texte zu erzeugen. Dieser digitale Stumpfsinn hat das Programm Chat-GPT sowie nun den Nachfolger GPT-4 hervorgebracht. Doch all das verbraucht auch enorm viel Energie.

Für das sogenannte Machine Learning, wie es bei Chat-GPT zum Einsatz kommt, schalten Programmierer mathematische Operationen, genannt Neuronen, zu großen, mehrdimensionalen neuronalen Netzen zusammen. Besteht der Algorithmus aus vielen Schichten, spricht man auch vom Deep Learning. Nach mehreren Trainingsdurchgängen sind die Netze zu ungeahnten Leistungen fähig.

Auf jedes Modell, das online geht, kommen Hunderte, die zuvor verworfen wurden

„Die meisten der Deep-Learning-Modelle werden auf spezieller Hardware trainiert", sagt Raghavendra Selvan, der als Juniorprofessor am Computer Science Department der Universität Kopenhagen arbeitet. Kernelemente der Hardware sind GPUs, Grafikprozessoreinheiten, die vor allem mit Blick auf Computerspiele entwickelt wurden und als Stromschlucker bekannt sind.

Würden nur ein paar Fachleute mit Chat-GPT herumspielen, wäre das noch kein Problem. Im vergangenen November, als Chat-GPT vorgestellt wurde, gab es 153 000 Zugriffe auf das System. Im Februar wurden dem Bot schon mehr als eine Milliarde Fragen gestellt. Wenig verwunderlich, dass sich Microsoft kürzlich die Dienste der Entwickler Chat-GPTs, der ursprünglichen Non-Profit-Organisation Open AI, für zehn Milliarden US-Dollar sicherte. Auch die führenden Tech-Unternehmen Google und Meta drängen mit ähnlichen Produkten auf den Markt.

An der Universität Kopenhagen hat sich ein Forscherteam um Raghavendra Selvan gefragt, welche Dimension der Energieverbrauch von all dem wohl annimmt. Die Wissenschaftler entwickelten die Software Carbontracker, mit der Entwickler ihren eigenen Energiebedarf messen können. Jeder Machine-Learning-Versuch gibt Selvan nun eine Rückmeldung über den Stromverbrauch. „Ein Durchgang für sich alleine scheint ganz harmlos. Doch wenn ich mir nach etwa zwei Monaten am Ende meines Projektes eine Zusammenfassung ausgebe, dann ist das eine schockierende Zahl."

Anfangs lernen Machine-Learning-Algorithmen so gut wie nie auf die gewünschte Art und Weise. Um das zu erreichen, gilt es, Dutzende Parameter einzustellen. Die bestimmen, wie die riesigen Datensätze verarbeitet werden – darauf folgt der nächste Test per Brute Force Learning. „Man sieht nur die Spitze des Eisberges", sagt Selvan. „Für jedes Modell, was tatsächlich online geht oder genutzt wird, gibt es Hunderte, die davor verworfen wurden."

Was das bedeutet, lässt sich an Chat-GPT betrachten. Dessen Grundlage bildet der Machine-Learning-Algorithmus Generative Pretrained Transformes 3, genannt GPT-3. Aus den verfügbaren Quellen folgerten Selvan und Kollegen, dass allein der letzte erfolgreiche Brute-Force-Learning-Durchgang von GPT-3 in etwa 189 Megawattstunden an Energie verbraucht und nach dänischem Strommix 85 Tonnen CO_2 freigesetzt hat. Das entspricht etwa dem neunfachen jährlichen CO_2-Ausstoß pro Kopf in Deutschland.

„Machine-Learning-Entwicklung ist hochgradig iterativ. Das ist das Unangenehme daran", sagt Selvan. Er ist sich sicher, dass Energieverbrauch und CO_2-Ausstoß, die zu Chat-GPT führten, noch um einiges höher sind. „Was wir ableiten konnten, ist nur ein Teil des tatsächlichen Verbrauchs in der Entwicklung von GPT-3." Dessen Algorithmus umfasst 175 Milliarden Neuronen, mehr als doppelt so viel wie ein menschliches Gehirn an Nervenzellen besitzt. Der GPT-3-Vorgänger GPT-2 wurde 2019 mit bloß sechs Milliarden Neuronen veröffentlicht.

Die Geschwindigkeit der Neuronenvermehrung überflügelt selbst das Moore'sche Gesetz. Dieses beschreibt, dass sich die Rechenoperationen pro Sekunde, die Computer durchführen können, alle etwa zwanzig Monate verdoppeln. „Jedoch verzweifacht sich gerade alle drei bis vier Monate der Bedarf an Rechenoperationen", sagt Selvan. Das Neuronenwachstum des Machine Learning verläuft also schneller, als Rechenpower bereitgestellt werden kann.

Um das Problem zu umgehen, schließen die Entwickler immer mehr GPUs parallel. Allein GPT-3 wurde auf 14 000 Grafikkarten gleichzeitig getestet. Firmen und Forschungseinrichtungen halten Schritt, indem sie eine nie da gewesene Anzahl von GPUs in Supercomputer integrieren.

„Trotzdem machen Anwendungen des Maschinenlernens bei unseren Allzweckrechnern derzeit nur etwa zehn Prozent aus", sagt Mirko Cestari, Teamleiter der Abteilung für Hochleistungscomputing der Organisation Cineca, einem Zusammenschluss italienischer Bildungseinrichtungen für Hochleistungscomputer. Vor Kurzem brachte Cineca in Bologna Leonardo ans Netz, den viertschnellsten Supercomputer der Welt, mit 14 000 GPUs an Bord.

„Doch es wird auf jeden Fall ein Wachstum in den nächsten Jahren erwartet", sagt Cestari im Hinblick auf Machine-Learning-Anfragen. Mit seinen Grafikkarten ist Leonardo gerüstet, doch das hat einen Preis: Sieben Megawatt an Leistung zieht er, das entspricht einem ICE 4 mit acht Waggons.

In Bologna ist man sich dieses Problems bewusst und sucht nach Lösungen. „Wir sind derzeit in Diskussion mit der Regionalregierung, ob sich die Abwärme nicht für Haushalte oder Industrie nutzen ließe", sagt Cestari. Die liegt schließlich bei gut 45 Grad Celsius. „Ideal zum Duschen", scherzt Cestari.

Auch auf der Softwareseite versuchen Fachleute, die neuronalen Netzwerke buchstäblich kleinzukriegen, etwa Wojciech Samek am Fraunhofer Heinrich-Hertz-Institut in Berlin. „Wie kann man die Anzahl der Operationen im Machine Learning reduzieren? Gibt es Redundanzen, die man nicht braucht?", fragt der Professor für maschinelles Lernen. Denn welche Rechenoperation exakt wie viel zum Endergebnis beiträgt, ist Forschern wie beim menschlichen Gehirn nicht vollends klar. Allerdings lässt sich das neuronale Netz der KI – anders als das biologische Denkorgan – digital sezieren.

„Vom Ergebnis ausgehend verteilen wir die Prozesse Schicht für Schicht mathematisch sinnvoll zurück", sagt Samek. Die Forscher schauen, wie stark die Neuronen einzelner Schichten im Deep-Learning-Algorithmus zum Ergebnis beigetragen haben. „Die Neuronen, die keine Relevanz haben, kann man dann entfernen."

Samek sieht noch viel mehr Möglichkeiten, Machine Learning effizienter zu gestalten. „In manchem Bereich wissen wir Menschen ja, was relevant ist", sagt Samek. Schon Kinder wissen, dass jedes Säugetier vier und nicht sieben Gliedmaßen hat. „Die Maschine weiß das nicht. Wenn es uns gelingt, dieses A-Priori-Wissen einzubringen, kann man viel effizienter trainieren und damit Energie sparen."

Forscher erwägen, die Abwärme von Supercomputern für Haushalte zu nutzen

Auch die Programmiersprache, die den Lern-Algorithmen zugrunde liegt, scheint Raum zum Energiesparen zu bieten. Neuronale Netzwerke werden überwiegend in Python geschrieben, diese Codiersprache soll 70 Mal energiehungriger sein als beispielsweise C++. Das liegt daran, dass Python für Menschen vergleichsweise einfach zu verstehende Befehle hat. Der Nachteil: Dies für Schaltkreise in Nullen und Einsen zu übersetzen, ist aufwendig. C++ hingegen ist aufwendig zu coden, doch für die Maschine recht schnell verständlich zu machen. „Python erlaubt Zugänglichkeit", sagt Raghavendra Selvan. „Hätten wir statt mit Python in C++ gecoded, hätten wir niemals diese Fortschritte beim Machine Learning gesehen." Generell wären neuronale Netze wohl nie so bedeutsam geworden, hätten Forscher über Energieeffizienz nachgedacht, sagt Selvan. „Doch jetzt müssen wir es tun."

Die Techriesen sind anderer Meinung. Google veröffentlichte kürzlich eine Studie, nach der der CO_2-Fußabdruck des Machine Learning nach einem weiteren Anstieg absinken und sich danach stabilisieren soll. Selvan bezweifelt die Aussagekraft dieser Untersuchung, sie sei zu beschränkt. Zudem wüchsen auch die Nutzerzahlen neuronaler Netze explosionsartig. Das führt zu einer weiteren offenen Frage: Niemand außer den Firmen weiß, was eine Textanfrage an die austrainierten Modelle eigentlich an Strom frisst. Für die SZ hat Selvan nun versucht, den Energiehunger zu beziffern. Dazu bedient er sich des Modells GPT-2 und skaliert die gewonnenen Zahlen entsprechend verfügbarer Daten für andere Sprachmodelle. Demnach verbraucht eine einzelne Anfrage von etwa 230 Wörtern 581 Wattstunden. Die eine Milliarde Anfragen an Chat-GPT im Februar hätten demnach einen Verbrauch von 581 Gigawattstunden verursacht, also in etwa den Stromverbrauch aller 170 000 Einwohner von Oldenburg pro Jahr. Das entspricht rund 244 000 Tonnen CO_2. Auch wenn neuere KIs auf effizienteren Maschinen liefen, sagt Selvan, so dürfte der Energiehunger von GPT-4 noch eine Dimension größer ausfallen.