

Schlimme Stromfresser

Das Training von Sprachmodellen wie ChatGPT ist enorm stromhungrig. Salzburger Forschende suchen nach einem Ausweg, revolutionäre Computerchips bieten sich dafür ebenso an wie eine intelligentere Verteilung der Rechenaufgaben.

Raimund Lang



Datenzentren sind leistungsstark, für komplexe Echtzeitberechnungen dauert die Datenübertragung von Geräten zur Cloud aber zu lange.

Künstliche Intelligenz (KI) ist nicht nur sprichwörtlich in aller Munde, sondern bindet auch real immer mehr Ressourcen in Rechenzentren. Grund dafür ist die hohe Rechenleistung, die von den stetig komplexer werdenden KI-Modellen benötigt wird. Mehr Rechenpower, ob nun in Form von CPUs oder von Grafikprozessoren (GPU), benötigt mehr Energie – und verursacht damit auch höhere CO₂-Emissionen. So verbrauchte allein das Training des KI-Modells GPT-3, das unter anderem Grundlage für den Chatbot ChatGPT war, mehr als 900.000 Kilowattstunden Strom. Das entspricht dem Jahresbedarf von knapp 300 durchschnittlichen Zwei-Personen-Haushalten in Österreich. Unternehmen und Forschungseinrichtungen versuchen deshalb auf verschiedene Weisen, diesem Problem beizukommen. So hat beispielsweise Amazon Web Services kürzlich um ungerechnet knapp 600 Millionen Euro ein Rechenzentrum samt angeschlossener Atomkraftwerke erworben.

Cloud oder Edge?

Im europäischen Projekt Energy-Efficient Large-Scale Artificial Intelligence for Sustainable Data Centers (Escade) wiederum arbeiten sieben Hauptpartner unter der Leitung des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI) an intelligenten Methoden der Lastverteilung zwischen der Cloud, also typischerweise einem Rechenzentrum, und der sogenannten Edge, also dem Bereich, in dem die zu bearbeitenden Daten erzeugt werden, wie auf Geräten und

Sensoren. Eine Schlüsselrolle kommt dabei dem Einsatz neuartiger Chipstechnologien zu.

Vereinfacht gesagt geht es also darum, ob eine Smartphone-App mit dem im Handy verbauten Prozessor arbeitet oder diese Rechenleistung von einem Serverzentrum durchgeführt wird. Dazu müssen die Daten natürlich über das Internet erst dorthin geschickt und dann wieder aus der sogenannten Cloud zurück auf das Gerät transferiert werden.

Das dreijährige Projekt startete im Mai 2023 und verfügt über ein Budget von rund fünf Millionen Euro. Gefördert wird es vom deutschen Bundesministerium für Wirtschaft und Klimaschutz (BMWK) und dem österreichischen Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK) über das Programm Green Tech & Tech for Green der FFG.

„Nicht alles wird in der Zukunft in der Cloud laufen“, sagt Jia Lei Du, Senior Researcher bei Salzburg Research, dem einzigen österreichischen Konsortialpartner bei Escade. „Bei großen KI-Anwendungen wie zum Beispiel automatisiertem Fahren oder mobiler Robotik wird man viele Aufgaben direkt vor Ort bei den Geräten ausführen lassen.“ Eine der Aufgaben der Salzburger bei Escade ist es, ein Modell zu entwickeln, das entscheiden kann, ob ein Rechenprozess besser in der Cloud oder in der Edge ausgeführt werden sollte.

„Besser“ ist dabei stets als Kompromiss aus verschiedenen Anforderungen zu verstehen. So erfolgen Berechnungen in der Edge zwar meist schneller als jene in der Cloud, weil die Datenübertragung zu und von Letzterer ent-

fällt. Allerdings sind Edge-Geräte in der Regel wenig leistungsstark, was ihre Anwendungsmöglichkeiten beschränkt. In der Cloud dagegen ist Rechenleistung zwar kein Problem, doch dafür kostet es Zeit, Daten hochzuladen, zu verarbeiten und wieder dorthin zurückzuschicken, wo das Ergebnis benötigt wird.

Beim autonomen Fahren beispielsweise gibt es Entscheidungen, die in Echtzeit getroffen werden müssen. Etwa dann, wenn ein unerwartetes Hindernis auf der Straße auftaucht. Sensordaten vom Auto an ein Rechenzentrum in der Cloud zu übertragen und auf das Berechnungsergebnis zu warten könnte in so einer Situation fatale Folgen haben. Als eine zusätzliche Stellgröße fließt auch der Energieverbrauch in diese Überlegungen ein. Schließlich verbraucht auch die Datenübertragung in die Cloud über mobiles Internet, etwa mittels 5G-Technologie, Energie.

Menschliches Gehirn als Vorbild

Als weiteren Einflussfaktor auf die Entscheidung, ob Cloud oder Edge der ideale Rechenort ist, untersuchen die Forscher neuartige Hardware. Insbesondere sogenannte neuromorphe Chips stehen hierbei im Zentrum des Interesses. Deren Architektur ahmt die extrem dichte Vernetzung der Nervenzellen im menschlichen Gehirn nach, das hohe Rechenleistung mit hoher Energieeffizienz kombiniert.

„Es gibt neuromorphe Chips, die in Tests um den Faktor zehn bis 100 energieeffizienter waren als aktuelle Hardware“, sagt Jia Lei Du. Die Projektpartner erhoffen sich durch die neue Hardware Effizienzsteigerungen für KI-

Modelle von bis zu 50 Prozent beim Training und bis zu 80 Prozent im produktiven Betrieb. In Escade wird anhand einer beispielhaften Bilderkennungsanwendung in einem Stahlwerk ein Vergleich des Energieverbrauchs klassischer Chipstechnologie mit neuromorphen Chips gezogen.

Die Anwendung beinhaltet eine Drohne, die über dem Schrottplatz des Stahlwerks kreist und den Bestand an Schrott filmt. Eine KI klassifiziert dabei den Schrott, ein wesentliches Ausgangsmaterial bei der Stahlherstellung, nach dessen metallurgischer Zusammensetzung. Anhand dieser Aufgabe soll sich zeigen, wie gut neuromorphe Chips im realen Einsatz performen. Die dabei gemessenen Kennwerte werden als Parameter in das Modell der Salzburger Forscher einfließen.

In einem ersten Schritt sollen die Ergebnisse genutzt werden, um im Vorfeld großer KI-Anwendungen festzulegen, wo welche Rechenaufgaben abgearbeitet werden. Fernziel ist es jedoch, dass Rechenaufgaben hochdynamisch und automatisiert zum jeweiligen Zeitpunkt der Berechnung zwischen Cloud und Edge verschoben werden. Edge und Cloud werden dann keine streng getrennten Einheiten mehr sein, sondern ineinander übergehen. Man spricht deshalb schon heute vom „Edge-Cloud-Kontinuum“.

„Es geht darum, Teilaufgaben von KI-Anwendungen innerhalb dieses Kontinuums optimal zu platzieren“, erklärt Jia Lei Du. „Also Algorithmen dort ausführen zu lassen, wo Rahmenbedingungen wie Geschwindigkeit oder Energieeffizienz gerade am besten geeignet sind.“

FORSCHUNG SPEZIAL ist eine entgeltliche Einschaltung in Form einer Medienkooperation mit österreichischen Forschungsinstitutionen. Die redaktionelle Verantwortung liegt beim STANDARD.

