Sustainability concerns with GenAI

Motivation

Wurde vor einigen Jahren nach künstlicher Intelligenz gefragt, gingen die ersten Gedanken wohl in Richtung Smart Assistants wie Amazon Alexa oder Apple Siri. Mittlerweile ist ChatGPT der Inbegriff für KI geworden.

Chatbots können dabei helfen die eigene Produktivität zu steigern. Nicht nur in der Recherche oder organisatorischen sowie kreativen Tätigkeiten wird KI eingesetzt, auch in der Wirtschaft wird sie immer häufiger genutzt. Unter anderem um Produktivität und Effizienz zu steigern und Zeitersparnisse zu erzielen.

KI-Modelle benötigen eine immense Menge an Daten, um Impulse korrekt verarbeiten zu können und dementsprechende Outputs zu liefern. Modelle werden einerseits mit menschlicher Unterstützung, andererseits selbstüberwacht trainiert. Außerdem ist ein regelmäßiges Aktualisieren der Trainingsdaten erforderlich. Das Training der Modelle ist jedoch nur ein Teil des Ressourcenverbrauchs, es kommen noch sämtliche Nutzeranfragen hinzu. Mit Ressourcen ist die gesamte Infrastruktur von Rechenzentren gemeint, also nicht nur die Computer selbst, sondern auch die Kühlungssysteme bzw. die Klimaanlage.

Der Stromverbrauch von Rechenzentren weltweit betrug im Jahr 2022 zwischen 240 und 340 TWh Strom. Dies entspricht 1 bis 1,3% des globalen Stromverbrauchs.

Definition:

Definition ist die künstliche bzw. maschinelle Nachbildung von menschlicher Intelligenz. Durch Training wird dem System eine Art Intelligenz verliehen. Wie genau die Entscheidung zustande kommt, ist nicht nachvollziehbar, ähnlich wie beim Denkprozess im Gehirn. (Black Box)

Teilgebiete der Künstlichen Intelligenz:

- Machine Learning
    - Supervised Learning
    - Unsupervised Learning
    - Reinforcement Learning
- (Deep Learning)
- Natural Language Processing

LLMs:

KI-Modelle, die Eingabe in natürlicher Sprache (Prompts) erfassen und verarbeiten und eine generierte Ausgabe in ebenfalls natürlicher Sprache zurückgeben. Vielen LLMs gelingen mittlerweile Aufgaben auf Menschenniveau auszuführen. LLMs gehören zur Gruppe der Pretrained Foundation Models. PFMs sind entweder auf

NLP spezialisiert (LLMs) oder auf grafische Verarbeitung. Aufgrund der Komplexität verbraucht das Training und die Nutzung von LLMs einen enormen Ressourcenverbrauch, was in entsprechenden umwelttechnischen und finanziellen Kosten mündet.

Dimensionen der Nachhaltigkeit:

Nachhaltigkeit bedeutet, dass die Bedürfnisse der aktuellen Generation so befriedigt werden, dass die Bedürfnisse zukünftiger Generationen nicht kleiner sein müssen.

Nachhaltigkeit umfasst ökologische, soziale und ökonomische Dimensionen. Ökologisch bedeutet, dass natürliche Ressourcen geschont und geschützt werden. Sozial bedeutet gerechte Lebensverhältnisse weltweit. Ökonomisch bedeutet langfristiger Wohlstand.

Co2-Fußabdruck

Der CO2-Fußabdruck beschreibt die Menge an Kohlenstoffdioxid, die durch den Menschen freigesetzt wird. Besonders rechenintensive Prozesse wie das Training großer KI-Modelle führen zu erheblichen Co2-Emissionen.

Maßnahmen umfassen auf individueller Ebene bspw. Umstieg auf öffentliche Verkehrsmittel, Verzicht auf Inlandsflüge, Nutzung erneuerbarer Energien und im Technologiebereich vor allem energieeffiziente Hardware, nachhaltige Trainingsmethoden, der Einsatz grüner Rechenzentren.

Neben den Emissionen während der Nutzung eines Produkts, gibt es auch weniger sichtbare, mit trotzdem meist größerem Einfluss.

Embodied Emissions (EE): fallen bereits während der Herstellung und Errichtung an, bzgl. KI beispielsweise bei der Produktion von Hardware oder dem Bau der Rechenzentren.

Operational Emissions (OE): werden während des Betriebs freigesetzt. Bei KI also bspw. Emissionen durch das Kühlen der Rechenzentren oder die Versorgung der Infrastruktur.

Stromverbrauch von KI

Der weltweite Stromverbrauch von Rechenzentren im Jahr 2022 lag bei etwa 230-340 TWh, was rund 1-1,3% des globalen Strombedarfs entspricht. Zusätzlich wird der Anstieg des Stromverbrauchs in den letzten Jahren um 20-40% aufgezeichnet. Rechenzentren verbrauchen nicht nur Strom, sondern auch etwa 1% der weltweiten Treibhausgasemissionen.

Eine regional fokussierte Betrachtung, auf die USA konzentriert, sieht den Anteil von Rechenzentrum am gesamten Stromverbrauch bei 2-3% mit einer prognostizierten Verdreifachung in den kommenden Jahren. Der Anteil von KI-Anwendungen macht

10-20% am Stromverbrauch der Rechenzentren aus. Dieser Wert könnte bis auf 70% steigen.

Verursachung der Emissionen:

EE:

- Herstellung der Hardware
- Training der Modelle
- Entwicklung der Modelle
- Bau von Rechenzentren
- Gewinnen der Ressourcen zur Hardwareherstellung

OE:

- Stromverbrauch der Server
- (Hosten der Modelle)
- (Verarbeitung der Anfragen)
- Kühlung

80-90% des von KI verbrauchten Stroms sind auf den Betrieb zurückzuführen.

Nur 32-40% des Stroms kommen aus erneuerbaren Energiequellen. Die Wahl des Standorts der Rechenzentren trägt zur Höhe der Emissionen bei. Je nach Region stehen unterschiedlich viele erneuerbare Energien zur Verfügung. Regionen mit viel Sonne, aber wenig Wind können nur tagsüber sauberen Strom bereitstellen. In den USA sind 95% aller Rechenzentren in Gebieten mit überdurchschnittlich CO2-Intensität. Auch die Wahl der Hardware in den Rechenzentren, sowie die Architektur der KI-Modelle und Trainingsstrategien machen einen Unterschied.

Auswirkungen

Es gibt Auswirkungen ökologischer, aber auch finanzieller Art. Der Informations- und Kommunikationssektor ist mittlerweile für 2% der globalen CO2-Emissionen verantwortlich. Eine andere Studie geht von 3,9% der globalen Treibhausgasemissionen aus. Bei Google sind die Treibhausgasemissionen seit 2019 um 48%, bei Microsoft seit 2020 um 29,1% gestiegen. Das fällt zeitlich mit dem KI Boom zusammen.

Konkrete Emissionswerte

Im Jahr 2918 haben US-Rechenzentren noch 31,5 Millionen Tonnen CO2 verursacht, von September 2023 bis August 2024 schon 105 Millionen Tonnen CO2.

Das Training von BLOOM, einem LLM mit 176 Milliarden Parametern, hat 50 Tonnen CO2 verursacht. Das Training von GPT-3 hat mehr als das Zehnfache, 552 Tonnen CO2 verursacht.

Bildgenerierung verbraucht so viel Strom wie eine Smartphoneladung.

Durch Rechenzentren werden in den nächsten Jahren so viele Treibhausgase wie 16 Millionen Autos ausgestoßen.

Herausforderungen

Während verbrauchte Energie einfach gemessen werden kann, sind Emissionen durch den Abbau der Ressourcen und die Hardwareherstellung selbst kaum erfassbar.

Eine weitere Herausforderung wird es, effizientere Hardware und Algorithmen zu entwickeln, ohne dabei in den Rebound-Effekt zu laufen, das heißt die Einsparungen durch intensivere Verwendung zunichtezumachen.

Es ist kaum möglich Rechenzentren durchgängig emissionsfrei zu betreiben, da erneuerbare Energien nichtkonstant verfügbar sind.

Ziele von Unternehmen und Staaten

Meta, Microsoft und Google verfehlen infolge der rasant fortschreitenden Entwicklung von KI ihre selbst gesetzten Nachhaltigkeitsziele. Grundsätzlich streben sie Klimaneutralität an, doch aufgrund des steigenden Energiebedarfs von KI-Anwendungen wird dieses Ziel verschoben oder vollständig aufgegeben. Außerdem schätzen 64% der Unternehmen den tatsächlichen Energieverbrauch durch KI als zu komplex für eine Messung ein.

Google strebt weiterhin Klimaneutralität bis 2030 an, doch ihre CO2-Emissionen sind von 2019 bis 2023 um 50% angestiegen. Die Klimaziele für 2024 kann Google aufgrund der Entwicklung von KI nicht einhalten.

Auf globaler Ebene sehen sich die Vereinigten Staaten, China sowie supranationale Akteure wie die EU mit vergleichbaren Herausforderungen konfrontiert.

Reduzierung von Emissionen

Um die durch KI verursachte Emissionen zu reduzieren, muss der Ressourcenverbrauch gesenkt werden. Einerseits kann die Nutzung von KI verringert werden, andererseits wäre ein effizienterer und gezielterer Umgang mit Ressourcen denkbar. Der erste Ansatz wird angesichts des Potenzials von KI als nicht praktikabel gesehen.

Entwicklungsansätze:

Sustainability by Design:

- Produkte und Dienstleistungen bereits so entwickeln, dass sie den Nachhaltigkeitsstandards genügen
- Kann auf technischer als auch organisatorischer Art erfolgen

- Langlebigkeit, einfache Reparierbarkeit, Wieder- bzw. Weiterverwendbarkeit, effiziente Energieverwendung

Die drei R der Kreislaufwirtschaft:

- Reduzierung: gleiche Rechenleistung mit weniger Hardwareressourcen
- Reusing: durch komponentenbasierte Systeme einzelne Bestandteile der gesamten Einheit wechselbar, bei Defekten, Upgrades
- Recycling: Hardware ein zweites Leben ermöglichen

Early Stopping Mechanismus:

- Bricht Training vorzeitig ab, um Training energieeffizienter zu gestalten
- Misst Leistungsmetriken und erkennt, wenn keine Verbesserungen mehr festgestellt werden, oder Genauigkeit sich sogar verschlechtert (Overfitting)
- Reduziert Trainingszeit
- Keine relevanten Unterschiede bei der Genauigkeit

KI gegen Emissionen

KI kann so eingesetzt werden, dass Emissionen reduziert oder gar vermieden werden. Es gibt zahlreiche Anwendungsfälle, in denen KI einen größeren Effekt erzielen kann. Durch KI verbessert sich der Zustand der Ökosysteme tatsächlich, wenn sie richtig eingesetzt wird.

Beispiele:

- Integration am Strommarkt
- Bauen oder Weiterentwicklung von Smart Buildings in Verbindung mit Internet Of Things zur Steigerung der Energieeffizienz
- Stadtentwicklung, intelligente Verkehrsnetze und Mobilität
- Tracken von Emissionen und anderen Umweltproblemen

Das Problem mit den Lösungen ist, dass sie teuer und komplex in der Anschaffung und somit auch noch nicht so weit verbreitet sind.

Komische Studie

Eine Studie untersucht, ob KI bei manchen Aufgaben weniger CO2 verursacht als Menschen. Dazu wurden vier KI-Modelle genommen, die sowohl Texte als auch Bilder generieren sollten. Bei den Co2 Emissionen wurden die durch Training und die durch Abfragen berücksichtigt. Bei den menschlichen Emissionen wurden Durchschnittswerte verwendet, wie viel CO2 pro Stunde und Bürger verursacht werden. Das Ergebnis waren 130- bis 1500-mal weniger CO2 zur Textgenerierung und 310- bis 2900-mal weniger CO2 zur Bildgenerierung. Die großen Intervalle entstehen durch den unterschiedlichen Verbrauch der Modelle, aber auch da ein US-Bürger im Durchschnitt einen größeren CO2-Fußabdruck hat als ein indischer Bürger.

Diese Erkenntnis lässt sich nicht generalisieren, da der Anwendungsfall sehr eingeschränkt ist.

Fazit

Den größten Teil des Stromverbrauchs durch KI macht Server-Hardware aus. Auch der Betrieb der Kühlungs-/Klimaanlagen trägt einen erheblichen Teil dazu bei.

Derzeit haben Rechenzentren einen Anteil von 1,0 bis 1,3 % am globalen Stromverbrauch, und es ist in den nächsten Jahren mit einem deutlichen Anstieg zu rechnen.

Es wird in Embodied und Operational Emissions unterschieden. Zu den EE gehören Emissionen durch die Förderung der Ressourcen für die Hardware, deren Herstellung und durch das Training der KI. Die OE setzen sich aus den Emissionen vom Strom zum Betrieb der KI-Server zusammen.

Das Bewusstsein für die hohen Emissionen durch KI ist bei Unternehmen und Staaten vorhanden. Jedoch mangelt es noch an der Umsetzung.