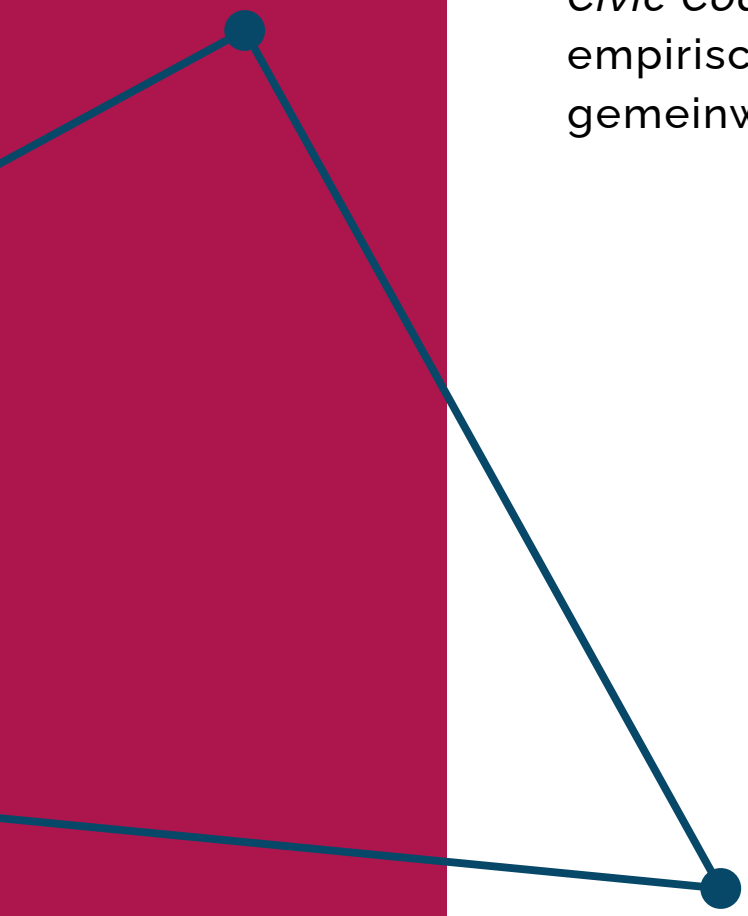
EXKURS

Theresa Züger, Judith Faßbender, Freia Kuper,
Sami Nenno, Anna Katzy-Reinshagen und Irina Kühnlein

Nachhaltigkeit von KI

Civic Coding – Grundlagen und empirische Einblicke zur Unterstützung gemeinwohlorientierter KI



CIVIC CODING
INNOVATIONSNETZ KI FÜR DAS GEMEINWOHL

Alexander von Humboldt
INSTITUT FÜR INTERNET
UND GESELLSCHAFT

Abbildungsverzeichnis

Tabellenverzeichnis

Exkurs: **Nachhaltigkeit** von KI

Forschungsliteratur zu Künstlicher Intelligenz (KI) unterscheidet zwischen *KI für Nachhaltigkeit* und *nachhaltiger KI* (van Wynsberghe 2021). Zu *KI für Nachhaltigkeit*, also dem Entwickeln und Einsetzen von KI-Anwendungen für das Erreichen von Nachhaltigkeitszielen, stellen beispielsweise Rolnick et al. (2022) verschiedene Einsatzmöglichkeiten lernender Systeme, die von der Optimierung CO2-armer Energiequellen bis zur Überwachung von Wäldern reichen, vor.

Bei *nachhaltiger KI* hingegen geht es darum, die KI-Anwendung selbst nachhaltig zu gestalten. In den letzten Jahren ist der hohe Energieverbrauch mancher KI-Anwendungen und damit auch die extremen CO2-Emissionen[1], die auf diese Anwendungen zurückzuführen sind, zunehmend in den Fokus gerückt.

Eine der am stärksten rezipierten Arbeiten zu den CO2-Emissionen von Machine-Learning-Modellen stammt von Strubell et al. (2019). Für diese Studie wurden Sprachmodelle trainiert und deren CO2-Ausstoß dokumentiert. Strubell et al. kommen zu dem Ergebnis, dass das Training insgesamt 284 Tonnen CO2 freisetzte. Dies entspricht ungefähr dem fünffachen Energieverbrauch eines Autos (inklusive Benzin) über seine gesamte Lebenszeit.

Ein weiteres oft genanntes Beispiel zur Emissionsanalyse stammt von Open AI.[2] Die Autor*innen haben untersucht, wie sich die Größe verschiedener Machine-Learning-Modelle in den letzten Jahren verändert hat. Sie kommen zu dem Schluss, dass sich der Rechenbedarf der Modelle zwischen 2012 und 2018 alle 3,4 Monate verdoppelt hat. Dieser Trend zu immer größeren Modellen wurde bereits an anderer Stelle festgestellt und besonders im Hinblick auf Nachhaltigkeit kritisch bewertet (siehe Bender et al. 2021). Da jene größeren Modelle mehr Energie benötigen, steigen auch die CO2-Emissionen. Entsprechend bedeutet die Verdoppelung des Rechenbedarfs alle 3,4 Monate, dass der Energiebedarf rasant steigt. Auch die Effizienzsteigerung von Computerchips in den vergangenen Jahren resultiert nur in einem partiellen Ausgleich. Die Autor*innen von Open AI merken dazu an, dass die Wachstumsrate von Machine-Learning-Modellen die Effizienzrate von Computerchips mittlerweile übertroffen hat.

Ein Großteil der Forschungsliteratur setzt sich mit den CO2-Emissionen auseinander, die beim Training eines Modells zustandekommen. Das ist nicht verwunderlich, denn hierbei handelt es sich um einen klar abgrenzbaren Zeitraum, in dem große Mengen Strom verbraucht werden. Der Energieverbrauch anderer Aspekte, wie der der Herstellung der Hardware oder der des eigentlichen Einsatzes des Modells, ist schwerer abzugrenzen und zu ermitteln.[3] In Bezug auf beide Bereiche lässt sich ein relevanter und zunehmend problematischer Energie- und Ressourcenverbrauch vermuten, der aufgrund eines Mangels an Transparenz und Dokumentation nicht gut belegbar ist. Allerdings gibt es auch Indizien dafür, dass das Training nicht immer der größte Treiber von Emissionen ist.

Training vs. Inferenz

In der Literatur zu maschinellem Lernen wird in der Regel zwischen *Training* und *Inferenz* unterschieden (Kaack et al. 2022). Beim Training handelt es sich um den Prozess, ein Modell an einen Datensatz anzupassen und es dadurch für eine bestimmte Aufgabe, wie Bild- oder Spracherkennung, zu optimieren. Mit Inferenz oder Inferenz-

1 Im Folgenden verwenden wir CO2 auch für CO2-Äquivalente.

2 https://openai.com/blog/ai-and-compute/

3 Kaack et al. gehen 2022 stärker auf diesen Aspekt ein. Wir werden im Folgenden darauf zurückkommen.

phase wird der eigentliche Einsatz des Modells beschrieben, um beispielsweise ein Bild zu erkennen. Wenn beispielsweise ein Satz bei Google Translate übersetzt wird, dann handelt es sich um eine Inferenz.

Eine Inferenz verbraucht üblicherweise im Vergleich zum Training relativ wenig Energie. Handelt es sich aber um viel genutzte Anwendungen, wie Google Translate, das täglich mehr als 100 Milliarden Wörter übersetzt,[4] summiert sich der Energieverbrauch (siehe auch Kaack et al. 2022). So schätzt NVIDIA-CEO Jensen Huang, dass etwa 80–90 % des Machine Learning Workloads auf Inferenzen zurückzuführen sind.[5] Auch Amazon Web Services (AWS) berichtet, dass etwa 90 % der Machine-Learning-Kosten in den eigenen Datenzentren für die Inferenzphase anfallen (siehe auch Patterson et al. 2021).[6] Auch wenn dies natürlich keine Angabe über den Stromverbrauch oder CO2-Ausstoß ist, lässt sich hier dennoch das Verhältnis erahnen. Diese Berichte stimmen mit einem 2021 erschienenen Forschungspapier von Meta überein (Wu et al. 2021). Dort wird berichtet, dass Metas Energieverbrauch für Machine Learning zu 30 % auf Training und Experimente und zu 70 % auf Inferenz entfällt. Zudem schätzen die Autor*innen, dass die CO2-Emissionen für die Produktion der nötigen Hardware etwa der Hälfte der Emissionen des Machine-Learning-Einsatzes (Datenverarbeitung, Training, Inferenz) entsprechen. Daraus lässt sich zusammenfassen, dass ein großer Teil des Energieverbrauchs von Machine-Learning-Prozessen schon vor dem Training beginnt und auch nach dem Training weiter anfällt.

Nachhaltigkeit entlang des KI-Lebenszyklus

Ein Versuch, nachhaltige KI nicht nur anhand des Trainings zu klassifizieren, stammt von Kaack et al. (2022). In der Forschungsarbeit von Kaack et al. (2022) wird nachhaltige KI auf drei Ebenen analysiert, die über eine reine Klassifikation des Trainings hinausgehen. Auf einer ersten Ebene werden die direkten Auswirkungen des Rechenaufwandes betrachtet: der Verbrauch des Trainings und der Einsatz von Machine-Learning-Modellen, aber auch die Infrastruktur, wie zum Beispiel Datenzentren. Auf einer zweiten Ebene geht es um die unmittelbaren Auswirkungen des Einsatzes von KI: KI-Anwendungen können eingesetzt werden, um die Effizienz erneuerbarer Energien zu steigern, aber auch, um die Kosten fossiler Energieträger zu senken. Je nach Einsatzzweck ist die Anwendung mehr oder weniger nachhaltig. Auf der dritten Ebene werden übergeordnete, gesellschaftliche Auswirkungen auf die Gesellschaft und auch mögliche Rebound-Effekte des Einsatzes von KI in den Blick genommen. Denn der Einsatz von KI kann zu Rebound-Effekten führen, die wiederum schlecht im Hinblick auf Nachhaltigkeit sind.

Ebenso wie die Verengung von *nachhaltiger KI* auf die Nachhaltigkeit des Modelltrainings zu kurz greift, so greift auch das Verständnis von Nachhaltigkeit als ökologischer Nachhaltigkeit zu kurz. Rohde et al. (2021) argumentieren, dass neben der ökologischen Dimension auch die wirtschaftliche und soziale Dimension von Nachhaltigkeit berücksichtigt werden müssen.

Wer hat welchen Anteil?

Bisher wurde beschrieben, wie hoch die CO2-Emissionen großer Machine-Learning-Modelle sind und welche Phasen des Lebenszyklus beachtet werden müssen, um die Nachhaltigkeit der Anwendungen zu ermitteln. Es stellen sich aber noch weitere Fragen: Was für einen Anteil hat Machine Learning an den globalen CO2-Emissionen? Mit einer Einschätzung im Kontext globaler CO2-Emissionen kann erst ein Vergleich des KI-Sektors zu anderen Industrien hergestellt werden. In der bereits erwähnten Forschungsarbeit von Kaack et al. (2022) findet sich eine der wenigen Schätzungen des Anteils von Machine Learning an den globalen CO2-Emissionen. Die Autor*innen treffen die Annahme, dass die meisten Machine-Learning-Prozesse nicht über private Computer, sondern über Datenzentren abgewickelt werden. Damit können der Anteil von Machine Learning in Datenzentren und der Anteil von Datenzentren

4 https://blog.google/products/translate/ten-years-of-google-translate/
5 https://www.hpcwire.com/2019/03/19/aws-upgrades-its-gpu-backed-ai-inference-platform/
6 https://aws.amazon.com/de/blogs/aws/amazon-ec2-update-inf1-instances-with-aws-inferentia-chips-for-high-performance-cost-effective-inferencing/

am gesamten Informations- und Telekommunikationssektor abgeschätzt werden. Die Autor*innen schätzen den Anteil von Machine Learning an den globalen CO2-Emissionen auf *0,025-0,05 %*. Sie betonen allerdings, dass es sich hier um eine grobe Schätzung handelt und es einige Unsicherheiten gibt.

Eine weitere Frage, die bisher allerdings wenig Beachtung gefunden hat, ist, wer die Modelle mit hohen CO2-Emissionen entwickelt und nutzt bzw. ob kleinere KI-Projekte auch auf ähnliche Zahlen kommen. Schaut man beispielsweise auf die vorher genannten Untersuchungen von Open AI zurück, ist auffällig, dass fast alle der immer schneller wachsenden Modelle von Alphabet stammen.[7] In der Tat ist es kein Geheimnis, dass die Forschung zu KI in den letzten Jahren vor allem durch einige wenige Big-Tech-Unternehmen und Eliteuniversitäten (oder in Kooperation) vorangetrieben wurde.[8] Kleineren Forschungseinrichtungen mit weniger Mitteln ist es fast nicht mehr möglich, die KI-Forschung an vorderster Front mitzubestimmen. Der Grund für diese Entwicklung ist vorrangig finanzieller Natur: Modelle dieser Größe zu trainieren, ist sehr kostenintensiv. Strubell et al. (2019) geben beispielsweise an, dass die Kosten für Cloud Computing im Rahmen ihrer wissenschaftlichen Experimente mehr als 100.000 Dollar betragen können. So eine Summe ist für kleinere Forschungseinrichtungen oder zivilgesellschaftliche Projekte in der Regel nicht einfach oder überhaupt nicht zu stemmen. Noch größere Modelle, wie Alpha Go Zero von Deep Mind, veranschlagten allein für die Hardware Kosten von 25 Millionen US-Dollar (Gibney 2017).

Wenn die bisher genannten Zahlen aber nur für große KI-Modelle gelten, stellt sich die Frage, was für Emissionen bei KI-Projekten von zivilgesellschaftlichen Organisationen, kleineren Forschungseinrichtungen oder kleineren bzw. mittelständischen Unternehmen zu erwarten sind. Hier ist die Datenlage leider nicht zufriedenstellend. Eine Ausnahme bildet die Analyse, die Marcus Voß für den Bericht von Rohde et al. (2021) durchgeführt hat. Wir haben die Resultate des Berichts repliziert und aktualisiert (Abbildung 4).

Für die Analyse wurden sogenannte Model Cards von Huggingface ausgewertet. Bei Huggingface handelt es sich um eine Hostingplattform für KI-Modelle, vorrangig für sogenannte Transformer, also eine bestimmte Art neuronaler Netze. Vortrainierte Modelle (siehe Abschnitt „Energieeffiziente Architekturen" im folgenden Kapitel) können hier hochgeladen werden und sind dann für die Allgemeinheit kostenfrei und niedrigschwellig zugänglich. Huggingface ist sehr populär in der Machine-Learning-Community, weil es einen einfachen Zugang zu sehr guten Modellen bietet. Man kann davon ausgehen, dass viele kleinere Projekte mit Huggingface arbeiten, und das macht es interessant für unsere Studie. Darüber hinaus werden die Modelle mit Model Cards ausgestattet, also einer Dokumentation über bestimmte Kennwerte des Modells. In Fig. 1 haben wir alle dokumentierten CO2-Emissionen vergleichend visualisiert und kontextualisiert. Es ist auffällig, dass das Training der meisten Modelle in etwa dieselbe Menge CO2 freisetzt wie ein PKW über eine Distanz von einem Kilometer. Nur bei einigen Modellen übersteigt der CO2-Ausstoß des Trainings einen Vergleichswert für eine Stunde Streaming in 4K-Qualität. Die hier dokumentierten Emissionen liegen demnach fernab der alarmierenden Zahlen und Vergleichswerte von Strubell et al. (2019).

Die Ergebnisse der Analyse gehen leider mit diversen Einschränkungen einher. Zum einen deckt die Analyse nur etwa 400 der über 50.000 von Huggingface gehosteten Modelle ab – diejenigen, die mit einer Model Card versehen sind.[9] Zum Teil wird die geringe Anzahl der Dokumentationen dadurch erklärt, dass der Nachhaltigkeit von KI erst

7 In dem genannten Beitrag von Open AI wird die Größe der Modelle auf Grundlage der entsprechenden Forschungspapiere ermittelt. Wir haben uns die Forschungspapiere angeschaut und festgestellt, dass die Mehrheit (8 von 15) der Autor*innen entweder über Google, Google Research oder Deep Mind mit Alphabet verbunden ist. Aber auch die restlichen Arbeiten stammen überwiegend von Big-Tech-Unternehmen oder Eliteuniversitäten, wie der Oxford University.

8 https://www.nytimes.com/2019/09/26/technology/ai-computer-expense.html

9 https://huggingface.co/models?sort=downloads

Experimente mit Huggingface oder mit unseren eigenen Modellen. Entsprechend widmen wir uns im Folgenden der Frage, wie die Dokumentation von Machine Learning verbessert werden kann und welche Maßnahmen bereits bekannt sind, um den CO2-Ausstoß zu minimieren. Trotz dieser hilfreichen Maßnahmen reicht jedoch der ausschließliche Blick auf das Training der KI-Modelle nicht aus, um die Nachhaltigkeit von KI-Systemen zu beurteilen und zu verbessern. Es braucht eine umfassendere Perspektive, die alle Phasen des KI-Lebenszyklus in Betracht zieht und auch die gesellschaftlichen Auswirkungen der KI-Anwendungen berücksichtigt.

Maßnahmen zur Dokumentation und Reduzierung der CO2-Emissionen

Im wissenschaftlichen Diskurs um nachhaltige KI finden sich verschiedene Handlungsempfehlungen und Techniken, um die Energieeffizienz von KI zu maximieren. Dabei kann zwischen Maßnahmen der Dokumentation und der Reduzierung unterschieden werden. Im Folgenden stellen wir einige Überlegungen, Techniken und Tools zur Dokumentation und Reduzierung des CO2-Ausstoßes von Machine-Learning-Modellen vor.

Dokumentation

Ein oft genanntes Problem für nachhaltige KI ist die mangelnde Dokumentation. Häufig ist nicht klar, wie umweltschädlich die Nutzung von KI-Systemen tatsächlich ist, weil die Daten zu einer Beurteilung nicht vorliegen (Zielinski et al. 2022). Entsprechend ist es notwendig, die Nachhaltigkeit von KI-Modellen in der Forschung wie auch in der Praxis zu dokumentieren. Es sollte also nicht nur die *Genauigkeit* des Modells, sondern auch die *Effizienz* angegeben werden. Die Genauigkeit gibt die Rate an, mit der das Modell in seinen Prognosen richtig liegt.[11] Die Effizienz gibt die Rate an, mit der das Modell in seinen Prognosen richtig liegt, in Relation zum Rechenaufwand des Modells. Es ist nicht selten, dass riesige Modelle trainiert werden, die nur marginal bessere Ergebnisse als bereits existierende Modelle erzielen. Schaut man nur auf die Genauigkeit, dann lässt sich hier dennoch sagen, dass die neuen Modelle bessere Ergebnisse liefern. Bezieht man aber auch den Rechenaufwand mit ein, achtet man also auf die Effizienz und nicht nur auf die Genauigkeit, dann erzielt das neue Modell schlechtere Resultate, weil die minimalen Verbesserungen in keiner Relation zum gestiegenen Energieverbrauch stehen.

Wie wir mit Blick auf Fig. 1 gesehen haben, gibt es Versuche, den CO2-Ausstoß von Machine-Learning-Modellen zu dokumentieren. Dennoch steht die Entwicklung erst am Anfang. Wie bereits erwähnt, geben nur etwa 400 der insgesamt ungefähr 50.000 Modelle, die auf Huggingface zu finden sind, ihren CO2-Ausstoß an. In der Forschung sieht es vergleichbar aus. In einer Studie von 2019 untersuchten Schwartz et al. Publikationen von mehreren renommierten Machine-Learning-Konferenzen. Sie fanden, dass mit Abstand in den meisten Publikationen eine Verbesserung der Genauigkeit und nicht der Effizienz als Hauptbeitrag des Forschungspapiers genannt wurde. Als Konsequenz plädieren Schwartz et al. für eine aussagekräftigere Dokumentation. Sie argumentieren, dass es mehr *Grüne KI* braucht, also KI, bei der vor allem auf Effizienz-und nicht auf Genauigkeitssteigerung geachtet wird.

11 Der Begriff Genauigkeit (Accuracy) wird in verschiedenen Kontexten unterschiedlich verwendet. Zum einen handelt es sich um eine Metrik, die die Güte eines Modells bemisst. In dieser Verwendung ist Genauigkeit eine Metrik unter vielen anderen, wie zum Beispiel F1 oder Präzision. Zum anderen wird der Begriff Genauigkeit aber auch in seiner alltäglichen Bedeutung genutzt und bezeichnet im Allgemeinen die Güte eines Modells, völlig undifferenziert hinsichtlich der zugrunde gelegten Metrik. Wir verstehen den Begriff im Folgenden in seiner alltäglichen Bedeutung und vor allem in Kontrast zur Effizienz, bei der es sowohl um Genauigkeit als auch um den Energieaufwand des Modells geht.

Neben einer häufigeren Dokumentation braucht es allerdings auch klare Dokumentationsstandards. Diese zu entwickeln, wird eine zentrale politische Aufgabe der kommenden Jahre sein, da der Standard der Dokumentation darüber entscheidet, wie aussagekräftig und wie vergleichbar der Ressourcen- und Energieverbrauch im KI-Sektor sind. Ohne hier bereits konkrete Vorschläge für einen Dokumentationsstandard entwickeln zu können, möchten wir einige Hinweise geben, warum diese Frage keineswegs trivial ist und möglicherweise ein eigenes Forschungsvorhaben erfordert.

Im Fall von Huggingface haben wir gesehen, dass die CO2-Emissionen des Trainings der Modelle angegeben wurden, nicht aber der Ort des Trainings. Wie aus Fig. 3 hervorgeht, kann sich der CO2-Ausstoß je nach Energiemix um ein Vielfaches unterscheiden, obwohl es immer derselbe Stromverbrauch ist. Das liegt daran, dass in manchen Ländern oder Regionen mehr oder weniger erneuerbare Energien eingesetzt werden. Das heißt aber auch, dass die bloße Angabe der CO2-Emissionen kein hinreichender Dokumentationsstandard ist. Interessanterweise wird auf Huggingface explizit genannt, dass die geografische Lage angegeben werden soll.[12] Dennoch haben nur wenige der Befragten sowohl ihre CO2-Emissionen als auch den Trainingsstandort angegeben.

Zu der Frage, welche Angaben sich als aussagekräftige Daten zum CO2-Ausstoß eignen, ist es naheliegend, den Stromverbrauch zu wählen. Hier gibt es jedoch die Schwierigkeit, dass der Stromverbrauch je nach Hardware variiert. Benutzt man beispielsweise eine neuere GPU, wird diese vermutlich stromeffizienter sein als eine alte Version.

Eine Metrik, die häufig in der Literatur hervorgehoben wird, sind die sogenannten „Floating Point Operations (FLOPs)". Ein FLOP kann grob als eine einzelne Rechenoperation, eine Multiplikation oder Addition, verstanden werden. 4,0 + 3,0 = 7,0 wäre also ein FLOP. Da sich ein Großteil der Berechnungen von Machine-Learning-Modellen auf simple Additionen oder Multiplikationen reduzieren lässt, bieten sich FLOPs als Maßeinheit an. Ein Beispiel: ResNet-50 – ein Modell zur Bilderkennung – braucht 4 Milliarden FLOPs, um ein Bild zu klassifizieren, während ResNet-152 11 Milliarden FLOPs für dieselbe Aufgabe benötigt. Damit kann hardware- und lageunabhängig angegeben werden, dass ResNet-50 energieeffizienter ist. Allerdings kann es sinnvoll sein, zusätzlich zu den FLOPs auch noch den Strom- und CO2-Verbrauch anzugeben. Bei den FLOPs handelt es sich um ein abstraktes Maß, das für viele nicht intuitiv ins Verhältnis zu setzen ist. Um ein besseres Bewusstsein für die Problematik zu schaffen, sollte es auch Möglichkeiten geben, diesen abstrakten Wert zu kontextualisieren. Manche Dokumentationstools geben beispielsweise zusätzlich an, wie weit man mit einem Pkw mit demselben CO2-Ausstoß hätte fahren können. Darüber hinaus kann auch der Einsatzzweck des Modells dokumentiert werden, damit dessen Nachhaltigkeit nicht nur in Bezug aufs Training, sondern auch auf die Inferenz hin bewertet werden kann.

Die folgende Tabelle listet existierende Werkzeuge zur Dokumentation von Stromverbrauch und CO2-Emissionen auf. Es lassen sich zwei Arten von Werkzeugen unterscheiden: 1) Python Libraries, mit denen man durch ein paar Zeilen Code den eigenen CO2- und Stromverbrauch während des Trainings messen kann, 2) Browser Apps, in die man nach dem Training eingibt, wie lange man trainiert hat und welche Hardware benutzt wurde, und die darauf aufbauend CO2- und Stromverbrauch berechnen.

12 https://huggingface.co/docs/hub/models-cards-co2

Bezeichnung	Format	Beschreibung	Basierend auf	URL
Code Carbon	Python Library	Kann in den Code des AI-Programms eingefügt werden, um den Stromverbrauch und die CO2-Emissionen während des Trainings zu dokumentieren.	Lottick et al. 2019	codecarbon.io
Carbon Tracker	Python Library	Siehe oben.	Anthony et al. 2020	github.com/lfwa/carbontracker
Experiment Impact Tracker	Python Library	Siehe oben.	Henderson et al. 2020	github.com/Breakend/experiment-impact-tracker
ML Emission Calculator	Browser App	Durch Angabe der Trainingszeit und des Cloud-Providers kann nach dem Training die CO2-Emission errechnet werden.	Lacoste et al. 2019	mlco2.github.io/impact
Green Algorithms	Browser App	Siehe oben.	Lannelongue et al. 2020	green-algorithms.org
Cloud Carbon Footprint	Browser App	Dokumentiert CO2-Emissionen verschiedener Cloud-Provider und berechnet persönlichen Ausstoß.	-	cloudcarbonfootprint.org

Tabelle 1: Verschiedene Tools für die Dokumentation der CO2-Emissionen

Reduzierung

Für die Reduzierung der CO2-Emissionen stehen zahlreiche technische Möglichkeiten zur Verfügung. Manche helfen dabei, den Strombedarf während des Trainings zu reduzieren, einige haben auch positive Auswirkungen auf die Inferenzphase. Einige der Techniken haben niedrigschwellige Zugangsvoraussetzungen und sind relativ einfach umzusetzen. Andere wiederum setzen relativ gute Kenntnisse der Software- und Hardwareseite voraus oder verkomplizieren den Programmierprozess. Im Folgenden listen wir einige der meistgenannten Techniken auf und erklären sie kurz. Die Liste ist nicht erschöpfend.

Hyperparametertuning

Machine-Learning-Modelle haben zwei Arten von *Parametern: Parameter* (manchmal auch *Gewichte* genannt) und *Hyperparameter*. Bei den Parametern handelt es sich um (Listen aus) Zahlen. Sie sind das, was im Trainingsprozess optimiert wird. In der Regel beginnt das Training mit einem Modell aus zufällig gewählten Parametern und diese werden auf Grundlage des Datensatzes so angepasst, dass das Modell die richtigen Vorhersagen trifft.

Im Gegensatz zu den Parametern werden die Hyperparameter nicht automatisch optimiert, sondern manuell gewählt. Hyperparameter sind für verschiedene Einstellungen beim Training verantwortlich, zum Beispiel wird damit festgelegt, mit welcher Geschwindigkeit das Modell trainiert wird. Welche Hyperparameter die besten für das

jeweilige Modell sind, kann nicht im Vorhinein berechnet, sondern muss durch Experimente ermittelt werden. Und hier liegt das Problem in Bezug auf Nachhaltigkeit: Für jede Kombination von Hyperparametern muss ein Modell trainiert werden. Und je mehr Kombinationen man testet, desto mehr Modelle werden trainiert und entsprechend steigt der Stromverbrauch.

Die vorherrschende Methode, die besten Hyperparameter zu finden (Hyperparametertuning), ist die Rastersuche (Grid Search). Dabei werden alle möglichen Kombinationen aus Hyperparametern getestet.[13] Die Rastersuche wird häufig als wenig nachhaltige Methode bewertet (Lacoste et al. 2019), weil übermäßig viele Modelle trainiert werden, obwohl am Ende nur ein einziges seinen Weg in die Anwendung findet. Eine Technik, um das Machine-Learning-Training nachhaltiger zu gestalten, ist, das Hyperparametertuning nicht mittels Rastersuche, sondern mittels probabilistischer Methoden durchzuführen (Bergstra and Bengio 2012). Anstatt alle Kombinationen durchzutesten, ermitteln statistische Methoden auf „smarte" Weise die richtigen Hyperparameter. Zwar müssen auch bei probabilistischen Methoden weiterhin mehrere Modelle trainiert werden, allerdings weitaus weniger als bei der Rastersuche.

Obwohl der Einsatz probabilistischer Methoden – neben einer besseren Ökobilanz – häufig auch zu besseren Ergebnissen als die Rastersuche führt (Bergstra and Bengio 2012), gibt es dabei Schwierigkeiten. Zum einen braucht es die Expertise, um probabilistische Methoden einzusetzen, zum anderen gibt es für die Rastersuche eine größere Auswahl an Libraries, sodass die Methode nicht mehr selbst implementiert werden muss.

13 Streng genommen werden nicht alle Kombinationen getestet, weil es sich bei den Hyperparametern häufig um reelle Zahlen handelt und es deswegen eine unendliche Anzahl an Kombinationen gibt. In der Praxis wird eine Menge an Werten vorher festgelegt und im Anschluss getestet.

Energieeffiziente Architekturen

Eine weitere Möglichkeit, Energie beim Training einzusparen, ist der Einsatz von energieeffizienten Architekturen, also bestimmten Arten von Machine-Learning-Modellen. Dazu zählen zum einen Modelle aus dem klassischen Machine Learning (im Gegensatz zum Deep Learning), zum Beispiel die Lineare oder Logistische Regression, Entscheidungsbäume oder Support-VectorMaschinen. Wie auch aus unseren eigenen Experimenten in Abbildung 5 ersichtlich wird, haben diese Modellarchitekturen einen weitaus geringeren Stromverbrauch als die meisten neuronalen Netze. Das liegt daran, dass klassische Architekturen weitaus weniger Parameter haben als neuere Architekturen und entsprechend weniger Rechenaufwand anfällt. Es ist erwähnenswert, dass klassische Architekturen nicht zwangsläufig schlechtere Ergebnisse liefern als die neuesten Deep-Learning-Modelle. Zwar haben tiefe neuronale Netze in den letzten Jahren den Diskurs bestimmt, dennoch liefern klassische Modelle bei vielen Aufgaben weiterhin vergleichbare Ergebnisse.

Aber auch der Einsatz neuronaler Netze kann energieeffizient gestaltet werden. In den vergangenen Jahren hat sich das Transfer Learning (oft auch Finetuning genannt) etabliert. Dabei wird ein Modell – häufig ein Transformer oder Convolutional Neural Network – an einem sehr großen Datensatz vortrainiert. In einem zweiten Schritt wird das Modell an einem weiteren, aber weitaus kleineren Datensatz an eine jeweilige Aufgabe angepasst. Wesentlich ist, dass das vortrainierte Modell kopiert und damit mehrfach verwendet werden kann. Es muss also nur einmal an einem großen Datensatz trainiert werden, kann das so erworbene Wissen allerdings für eine potenziell unbegrenzte Anzahl neuer Aufgaben nutzen. Gerade im Kontext ökologischer Nachhaltigkeit kann man Transfer Learning in Analogie zu Recycling verstehen. In gewisser Weise wird das vortrainierte Modell wieder und wieder recycelt. Und weil es jedes Mal nur an einem kleinen Datensatz angepasst wird, ist der Energieverbrauch kleiner als jedes Mal von Grund auf neu zu trainieren.

Modellkomprimierung

Bei der Modellkomprimierung wird das Modell verkleinert und verursacht entsprechend weniger Rechenaufwand und hat damit einen niedrigeren Energieverbrauch. Für die Modellkomprimierung sind vor allem zwei Methoden populär: Quantisierung und Pruning.

Bei der Quantisierung wird die Komplexität der Parameter reduziert und so die Recheneffizienz optimiert.[14] Meistens handelt es sich bei den Parametern der Modelle um (Listen von) Zahlen, und zwar um Zahlen mit vielen Nachkommastellen. Werden die Nachkommastellen reduziert oder ganz beseitigt, wird weniger Rechenleistung benötigt. Natürlich büßt das Modell dadurch an Genauigkeit ein. Man kann das mit einer Matheaufgabe aus der Schule vergleichen: Wenn ich in der Hälfte der Rechnung mein Zwischenergebnis runde und dann damit weiterrechne, kann ich am Ende auf ein falsches Endergebnis kommen. Auch beim Pruning geht es darum, die Modellkomplexität zu reduzieren und damit die Recheneffizienz zu erhöhen. Beim Pruning werden einige der Parameter aber nicht nur verkleinert, sondern komplett entfernt. Bei beiden Methoden gilt es, einen guten Ausgleich zu finden. Die Komplexität der Parameter sollte so weit wie möglich reduziert werden, während die Genauigkeit des Modells so weit wie möglich unverändert bleiben sollte. Python Frameworks wie zum Beispiel PyTorch haben Techniken zur Quantisierung[15] und auch zum Pruning[16] schon implementiert. Für eine detailliertere Darstellung siehe Menghani 2021.

Anders als die bisher genannten Methoden hat Modellkomprimierung nicht nur Einfluss auf das Training, sondern auch auf die Inferenzphase. Kleinere Modelle verbrauchen auch im Einsatz weniger Energie. Gerade für Modelle, die häufig zum Einsatz kommen, lohnt sich die Komprimierung also.

14 Etwas präziser: Meist werden die Parameter von *32-bit floating-point* auf *8-bit fixed-point integers* reduziert.

15 https://pytorch.org/docs/stable/quantization.html

16 https://pytorch.org/tutorials/intermediate/pruning_tutorial.html

Geplante Trainingsphasen

Beim Training eines Machine-Learning-Modells handelt es sich um einen unterbrechbaren Prozess. Das heißt, dass es möglich ist, das Training in verschiedene Phasen aufzuteilen und zwischen diesen Phasen zu pausieren. Dass das Training nicht an einem Stück durchgeführt werden muss, ermöglicht es, die Trainingsphasen so zu timen, dass sie den besten Energiemix abpassen. Wie aus Fig. 3 weiter oben hervorgeht, variiert die Kohlenstoffintensität (gCO2/kWh) unter anderem mit der Tageszeit. Weil tagsüber die Sonne scheint, steht beispielsweise mehr umweltfreundliche Solarenergie zur Verfügung. Entsprechend erzeugt das Training zu diesen Zeiten weniger CO2. Und wenn das Training auf solche Phasen gelegt wird, erzielt es eine bessere CO2-Bilanz, als wenn einfach an einem Stück trainiert werden würde (für eine ausführliche Diskussion siehe Wiesner et al. 2021).

Es muss allerdings erwähnt werden, dass diese Technik mit praktischen Einschränkungen verbunden ist. Häufig wird das Training auf die Nacht oder das Wochenende gelegt, weil das Modell dann am nächsten Arbeitstag zur Verfügung steht. Außerdem dauert das Training natürlich länger, wenn es unterbrochen wird, und das ist häufig nicht mit existierenden Deadlines vereinbar. Ob in der Praxis beim Unterbrechen in mehrere Trainingsphasen die Vorteile die Nachteile aufwiegen, ist schwer im Allgemeinen festzulegen und sollte von Fall zu Fall beurteilt werden.

Hardwareauswahl

Durch den Einsatz erneuerbarer Energien kann der CO2-Ausstoß für Training und Inferenz erheblich gesenkt werden. Im Umkehrschluss heißt das allerdings, dass dem CO2-Ausstoß für die Herstellung der Hardware eine entsprechend höhere Relevanz zukommt. Man spricht hier von verkörperten Emissionen, also Emissionen, die nicht durch den Einsatz, sondern durch die Produktion der Hardware zustande gekommen sind (Wu et al. 2021, Gupta 2021). Auch bei der Hardwareauswahl gibt es einige Stellschrauben, die den CO2-Ausstoß verringern können.

Eine erste Überlegung ist, ob sich die Anschaffung eigener Hardware lohnt oder ob auf Cloud Computing zurückgegriffen werden soll. Neben einem normalen Computer benötigt das Training von Machine-Learning-Modellen in der Regel zusätzliche Grafikprozessoren (GPUs). Obwohl diese ursprünglich für die Computerspielindustrie hergestellt und optimiert wurden, eignen sich GPUs dazu, die Trainingszeit drastisch zu verringern. Kaack et al. 2022 berichten, dass die Treibhausgasemissionen für solche dezentrale Hardware zu 40–80 % von verkörperten Emissionen stammen. Im Fall von Datenzentren sind es nur 10 %. Die relevante Frage ist: Wie hoch ist die erwartete Auslastung der Hardware? Anders als bei Datenzentren besteht bei persönlicher Hardware wenig Auslastung. Nur wenn das Modell gerade trainiert wird, ist die GPU in Benutzung. Falls aber doch eigene Hardware angeschafft wird, können Nachhaltigkeitssiegel, wie der Blaue Engel, das Europäische Umweltzeichen, der Energy Star oder das TCO-Gütesiegel für Orientierung sorgen (Rohde et al. 2021).

Auch bei der Auswahl des Cloud-Computing-Anbieters gibt es Nachhaltigkeitskriterien zu berücksichtigen. Es gibt verschiedene Kennzahlen zur Bewertung der Nachhaltigkeit von Rechenzentren. Eine davon ist die Power Usage Effectiveness (PUE). Die PUE gibt an, wie viel des benötigten Stroms tatsächlich in das Rechenzentrum fließt und wieviel in den Überbau, wie beispielsweise für die Kühlung. Ein perfekter PUE ist 1. Dies bedeutet, dass 100 % des Stroms für das Rechenzentrum verwendet werden. Google, neben Microsoft Azure und Amazon Web Services der größte Anbieter von Cloud Computing, hat mit 1,1 eine ausgezeichnete PUE.[17] Der EU-Durchschnitt liegt bei 1,8 (Avgerinou et al. 2017). Neben der PUE sollten auch Carbon Usage Effectiveness (CUE) oder Water Usage Effectiveness (WUE) berücksichtigt werden (Rohde et al. 2021). Auch für Datenzentren gibt es einige Nachhaltigkeitsauszeichnungen, wie den Code of Conduct der EU[18] oder die Certification of Energy Efficiency for Data Centers (CEEDA).[19]

17 https://www.google.com/about/datacenters/efficiency/

18 https://joint-research-centre.ec.europa.eu/energy-efficiency/energy-efficiency-products/code-conduct-ict/code-conduct-energy-efficiency-data-centres_en

19 https://www.ceedacert.com/about-ceeda

Literaturverzeichnis

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020, July 6). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *ArXiv:2007.03051 [Cs, Eess, Stat]*. ICMLWorkshop on "Challenges in Deploying and monitoring Machine Learning Systems", http://arxiv.org/abs/2007.03051

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281–305

Gibney, E. (2017). Self-taught AI is best yet at strategy game Go. *Nature*. https://doi.org/10.1038/nature.2017.22858

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *ArXiv:2002.05651* [Cs]. http://arxiv.org/abs/2002.05651

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*. https://doi.org/10.1038/s41558-022-01377-7

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning. *ArXiv:1910.09700 [Cs]*. http://arxiv.org/abs/1910.09700

Lannelongue, L., Grealey, J., & Inouye, M. (2020). Green Algorithms: Quantifying the carbon footprint of computation. *ArXiv:2007.07610 [Cs]*. http://arxiv.org/abs/2007.07610

Lottick, K., Susai, S., Friedler, S. A., & Wilson, J. P. (2019). Energy Usage Reports: Environmental awareness as part of algorithmic accountability. *ArXiv:1911.08354 [Cs, Stat]*. http://arxiv.org/abs/1911.08354

Menghani, G. (2021). Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ArXiv:2106.08962 [Cs]*. http://arxiv.org/abs/2106.08962

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon Emissions and Large Neural Network Training. *ArXiv:2104.10350 [Cs]*. http://arxiv.org/abs/2104.10350

Rohde, F., Wagner, J., Reinhard, P., Petschow, U., Meyer, A., Voß, M., & Mollen, A. (2021). Nachhaltigkeitskriterien für künstliche Intelligenz. *Schriftenreihe Des IÖW*, 220, 21

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Bengio, Y. (2023). Tackling Climate Change with Machine Learning. *ACM Computing Surveys, 55*(2), 1–96. https://doi.org/10.1145/3485128

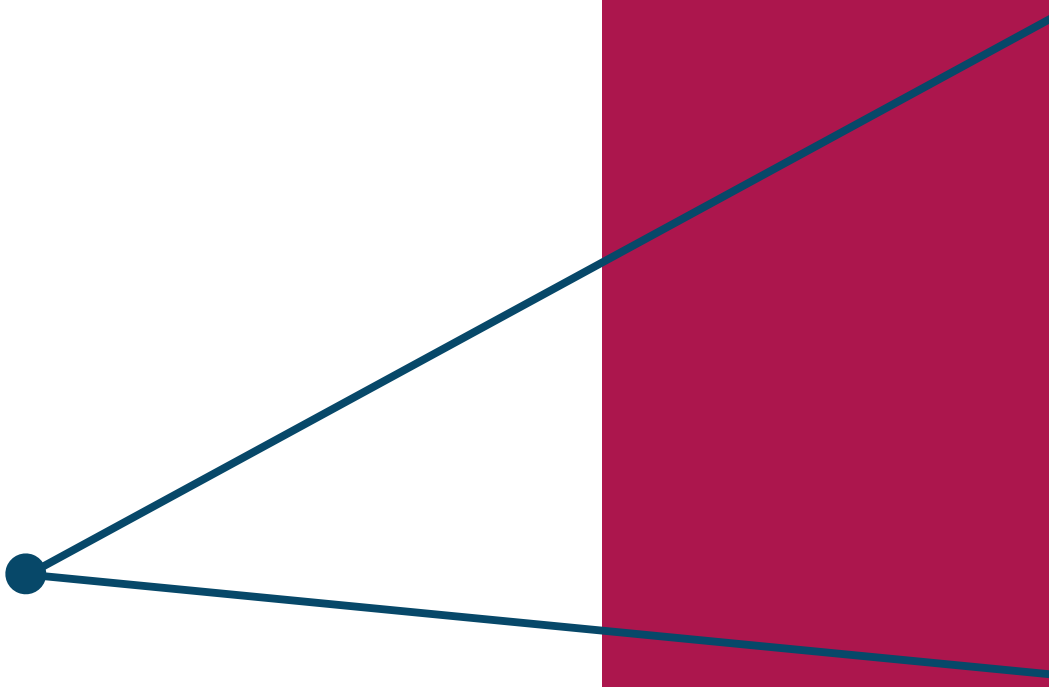
Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). Green AI. *ArXiv:1907.10597 [Cs, Stat]*. http://arxiv.org/abs/1907.10597

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., Hazelwood, K. (2021). Sustainable AI: *Environmental Implications, Challenges and Opportunities.*

Zielinski O., et. al (2022). Wege in eine ökologische Machine Economy, Hrsg. Daniel Wurm, https://www.germanwatch.org/de/85538.



Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

Bundesministerium
für Arbeit und Soziales

Bundesministerium
für Familie, Senioren, Frauen
und Jugend

civic-coding.de