Energiebedarf von digitalen Zukunftstechnologien und Anwendungen

Technologien, Herausforderungen, Lösungsansätze und politische Rahmenbedingungen

Roadmap Energieeffizienz 2050
2. Input-Papier für die AG Digitalisierung

1 Wodurch entstehen bei digitalen Zukunftstechnologien und Anwendungen die größten Energiebedarfe?

Die Optimierung von Energieverbräuchen und Prozessen kann einen erheblichen Beitrag zur Energieeffizienz und Energiebedarfsreduktion in allen Sektoren leisten. Gleichzeitig haben digitale Infrastrukturen und Zukunftstechnologien selbst teilweise einen erheblichen Energiebedarf. So kommt die Bitkom (2020) zu dem Schluss, dass im Jahr 2020 die Herstellung, der Betrieb und die Entsorgung digitaler Endgeräte und Infrastrukturen THG-Emissionen zwischen 1,8 und 3,2 Prozent der gesamten weltweiten THG-Emissionen verursachen.

Abbildung 1: Die Datenverarbeitungskette (eigene Darstellung)

Insbesondere die neuen digitalen Technologien treiben den Energiebedarf der IKT(-Infrastruktur) immer weiter voran. Die Energiebedarfe der unterschiedlichen digitalen Zukunftstechnologien und deren Anwendungen können entlang der Datenverarbeitungskette genauer betrachtet werden. Diese Kette ist in Abbildung 1 dargestellt. So entstehen Energiebedarfe insbesondere in den sechs Schritten: der Datenerfassung, der Datenübertragung, der Datenspeicherung, der Datenverarbeitung, der erneuten Datenübertragung und der Datenausführung. Je nach Technologie und Anwendung unterscheiden sich die Anteile des Energiebedarfs in den jeweiligen Schritten. Durch die Betrachtung des Energiebedarfs entlang dieser Datenverarbeitungskette ist es auch möglich differenziertere Aussagen darüber zu treffen in welchen Sektoren und wo die Bedarfe geographisch anfallen sowie damit einhergehende Energiebedarfsspitzen der unterschiedlichen Technologien und Anwendungen zu identifizieren.

Der Energiebedarf des ersten und letzten Schrittes der Kette, die Datenerfassung und die Datenausführung entsteht insbesondere in den Endgeräten, welche die Daten erfassen und am Ende der Kette mit den verarbeiteten Daten eine bestimmte Aktion ausführen. Der Energiebedarf des ersten und letzten Schrittes der Kette, die Datenerfassung und die Datenausführung entsteht insbesondere in den Endgeräten, welche die Daten erfassen und am Ende der Kette mit den verarbeiteten Daten eine bestimmte Aktion ausführen. Dahingegen entsteht der Energiebedarf der Datenübertragung insbesondere im Netzwerk und bei den Edge-Knoten (dies sind zum Beispiel Gateways oder Computer die zur Kommunikation mit anderen Knoten genutzt werden). Der Energiebedarf der Datenspeicherung und der Datenverarbeitung fällt sowohl in der Cloud als auch in Rechenzentren an.

Im Folgenden wird für zentrale digitale Technologien und Anwendungen betrachtet, wo die Hauptlast der Energiebedarfe anfällt. Im Vergleich zu aufwendigen Datenverarbeitungsprozessen benötigen die Datenerfassung und die Datenübertragung in der Regel deutlich weniger Energie, sind aber dennoch nicht zu vernachlässigen. Diese beiden Phasen sind insbesondere bei massiven Datenmengen relevant. Daher werden die digitalen Technologien, deren Energiebedarf insbesondere durch die hohen Datenmengen in den beiden Phasen der Datenerfassung und Datenübertragung entstehen, an den beiden Anwendungsbeispielen "autonomes Fahren" und "digitaler Zwilling" ausführlicher diskutiert.

1.1 Künstliche Intelligenz

Die Künstliche Intelligenz (KI) ist eine typische rechenorientierte Zukunftstechnologie. Ihr Prozess besteht aus drei Schritten: Datenvorverarbeitung, Modelltraining und Modellanwendung, von denen die beiden ersten energieintensiv sind und der zweite Schritt, das Modelltraining, noch deutlich energieintensiver ist als die Vorverarbeitung. Daher zentralisiert sich der Energiebedarf von KI technisch gesehen in den Rechenzentren, in denen die ersten beiden Schritte betrieben werden. Der Einsatz von bereits trainierten Modellen benötigt im Einzelnen im Endgerät dann nur noch verhältnismäßig wenig Energie. Wenn diese Modelle jedoch in der Masse eingesetzt werden, sind auch die kumulierten, dezentralen Energiebedarfe der Endgeräte nicht zu vernachlässigen (siehe auch weiter unten anhand des Beispiels des autonomen Fahrens). Geographisch gesehen fallen die Energiebedarfe für die Daten-Vorverarbeitung und das Modell-Training insbesondere dort an, wo die Rechenzentren liegen, während der Schritt der Datenausführung in der Maße dann dezentrale Energiebedarf hervorruft.

Im Folgenden werdend die beiden energieintensivsten Schritte der künstlichen Intelligenz genauer betrachtet.

Daten-Vorverarbeitung

Abgesehen von der Datenbereinigung arbeitet die Datenvorverarbeitung von KI hauptsächlich an der Datenanreicherung und Merkmalsextraktion. Es gibt unterschiedliche Verarbeitungsprozesse für Texte, Bilder, Sprache und Videos, die auf die verschiedenen Datentypen abzielen und unterschiedlich komplex sind. Die Verarbeitung von Texten stellt hierbei das einfachste Beispiel dar, da es sich bei Texten um eindimensionale Daten handelt. Die Verarbeitung von Bildern ist bereits etwas komplizierter, da sie zwei (schwarz/weiß) beziehungsweise drei (farbig) Dimensionen umfasst. Sprachesetzt sich aus dem Text plus einer Lautstärke zusammen. und Videos umfassen alle drei Komponenten, Text, Sprache und Bilder. Am einfachsten Beispiel, der Verarbeitung von geschriebenen Texten, wird deutlich, wie erheblich die Emissionen der Datenverarbeitung sein können. So kommen Strubel et al. (2019) zu dem Ergebnis, dass die Daten-Vorverarbeitung und Bereinigung einer normalen Natural Language Processing (NLP)-Pipeline für Texte bis zu 35 t CO_2 emittiert, um zuverlässige Ergebnisse zu erhalten (Strubell et al., 2019). Dies entspricht mehr als dem Vierfachen der jährlichen deutschen Pro-Kopf-Emissionen. Natural Language Processing kann auch für gesprochene Sprache eingesetzt werden, die Komplexität und die damit einhergehenden Emissionen erhöhen sich dadurch noch einmal deutlich.

Modell-Training

Das Training eines leistungsstarken maschinellen Lernalgorithmus bedeutet oft, dass riesige Computereinheiten tagelang, wenn nicht gar wochenlang, laufen müssen. Insbesondere die Feinabstimmung, die erforderlich ist, um einen Algorithmus zu perfektionieren, kann sehr rechen- und damit energieintensiv sein. Dies ist zum Beispiel dann der Fall, wenn verschiedene neuronale Netzarchitekturen durchsucht werden, um die beste Konfiguration zu finden. Von OpenAI veröffentlichte Daten zeigen, dass sich die benötigte Rechenleistung für konkrete Modelle in den letzten Jahren etwa alle 3,4 Monate verdoppelt hat. Dies entspricht in der Zeit von 2012 bis 2018 einem Faktor von 300.000 (vgl. Abbildung 2) (OpenAI, 2018). Diese erhöhte Rechenleistung ist darauf zurück zu führen, dass bei der Programmierung derzeit die Präzision und die Leistungsstärke der KI im Vordergrund stehen, nicht der Energiebedarf. Durch diese erhöhte Rechenleistung nimmt auch der Energiebedarf derzeit stetig weiter zu.

Das Team von UMass Amherst nahm NLP als Beispiel, um zu messen, wie viel Energie das Trainieren eines Modells verbraucht (Strubell et al., 2019). Ein einfaches maschinelles Übersetzungsprogramm, wie das von Wang analysierte, benötigt 27 kWh für den gesamten Trainingsprozess, der die durchschnittliche Stromverbrauchseffizienz des Rechenzentrums berücksichtigt. Um jedoch zuverlässige maschinelle Übersetzungen zu erzielen, verbraucht das neueste Modell Neural Architecture Search 656 MWh und emittiert im Durchschnitt 284 t CO_2, was etwa dem Fünffachen der Emissionen des Lebenszyklus eines Autos entspricht (57 t).

Anwendungen	Beispiel-Modelle	Energiebedarf pro Probe (J)	Energiebedarf für typische Datasets (kWh)
Maschinelle Übersetzung	Transformer (Übersetzungs-modell)	0,0098	0,098
Spracherkennung	Deep Speech 2	2,33	0,00061
Sprachmodellierung	Long Short-Term Memory	0.21	0.00245
Bilderkennung	ResNet 50	0.29	0.09667
Bilderkennung	Inception V3	0.45	0.15
Bilderkennung	VGG 16	0.48	0.16

Zusammenfassend kann gesagt werden, dass sich die Komplexität der Modelle, der Bedarf an enormen Datenmengen und die Vielfalt der Modellauswahl deutlich auf den Energiebedarf der KI auswirkt. Das Thema des enormen Energiebedarfs der KI gewinnt in den letzten Jahren zunehmend an Bedeutung, wodurch sich ein deutlicher Forschungsbedarf eröffnet.

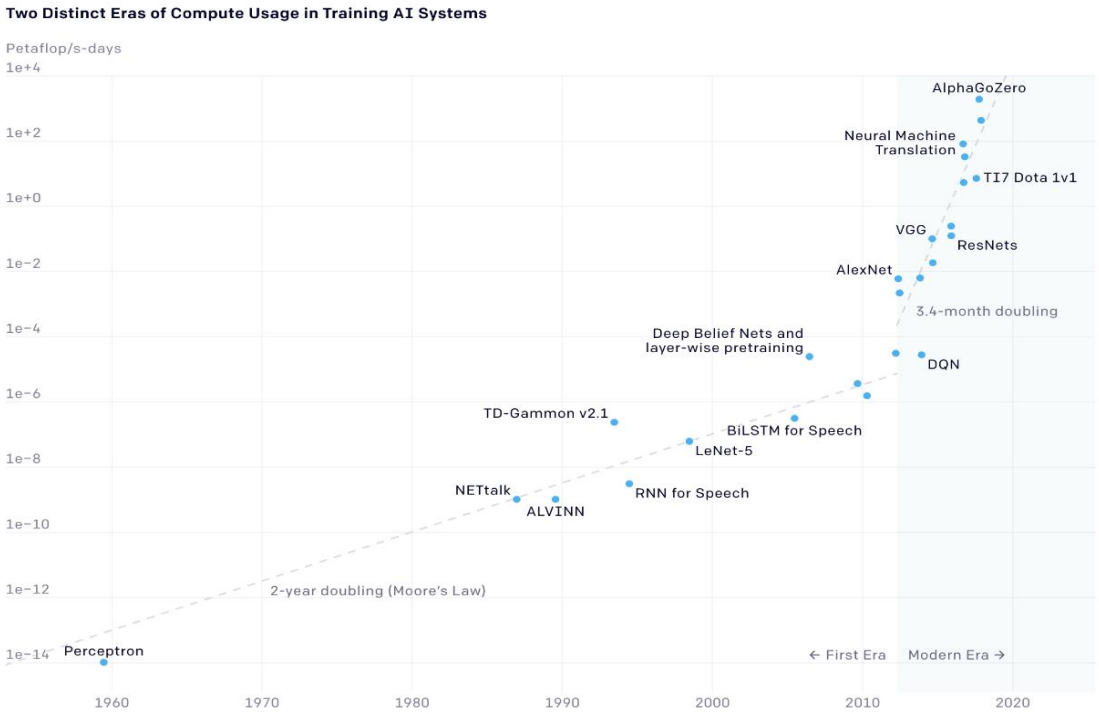

Abbildung 2: Die Entwicklung des Rechenbedarfs von KI-Trainingssystemen (OpenAI, 2018)

1.2 Virtuelle Realität

Ähnlich wie die KI ist auch die Virtual Reality (VR) eine typische rechenorientierte Zukunftstechnologie, deren Stromverbrauch zu etwa 80 Prozent aus Rechen- und Speicheranteilen besteht (Leng et al., 2019). Technisch gesehen dezentralisiert sich ihr Energiebedarf jedoch auf die Endgeräte, auf denen VR gerendert wird. Daher sollte mit zunehmendem VR-Einsatz auch der Energiebedarf der Datenübertragung zwischen Servern und Endgeräten ausführlicher betrachtet werden. Dies wird in Anwendungsbeispielen dargestellt.

Heutige VR-Systeme sind so konzipiert, dass gut etablierte Techniken, die für konventionelle Videos entwickelt wurden, wiederverwendet werden können. Diese Strategie beschleunigt den Einsatz von VR-Videos, verursacht aber einen erheblichen Energieaufwand (etwa 40 Prozent des gesamten Energiebedarfs), indem für jedes Einzelbild eine Sequenz von sphärisch-planarer projektiver Transformation berechnet wird. Der hierdurch entstehende Energiebedarf ist circa doppelt so hoch wie bei konventionellen Videos (Leng et al., 2019). Durch höhere Bildraten, eine höhere Auflösung und mehr Interaktionen könnte der Energiebedarf weiter steigen, insbesondere wenn der Trend weiter zu VR-Spielen geht. Zudem muss auch die Leistungsfähigkeit der Endgeräte stetig weiterwachsen, um die Daten entsprechend verarbeiten zu können.

Im Vergleich zu KI ist bei VR bei den Dienstleistungsanbietern bereits ein Umdenken bezüglich des Energiebedarfs eingetreten, da dieser in direktem Zusammenhang mit der Dienstleistungsqualität steht. Viele führende Unternehmen arbeiten daran, den Energiebedarf der VR weiter zu reduzieren.

1.3 Blockchain

Ähnlich wie bei VR verbraucht die Blockchain ihre Energie hauptsächlich durch dezentrale Berechnungen. Der Energieaufwand hängt hierbei jedoch eher von der Schwierigkeit der Kryptologie-Lösung als von der Datenmenge ab.

Die Blockchain-Technologie wird schon seit längerem wegen ihres enormen Energiebedarfs debattiert. Schätzungen zu Folge wird derzeit für eine einzige Transaktion mehr als 100 kWh Strom benötigt. Wie in Abbildung 3 am Beispiel der Blockchain-Währung Bitcoin dargestellt, steigt der geschätzte Energiebedarf von Bitcoin von knapp 4 TWh im Jahr 2016 auf etwa 60 bis 80 TWh pro Jahr im Juli 2019 (Cambridge, 2020). Laut der Schätzung von Cambridge liegt der derzeitige Energiebedarf (Mitte September 2020) bei etwa 66 TWh jährlich, was in der Größenordnung des jährlichen Stromverbrauchs von Ländern wie Österreich und Norwegen liegt (Sedlmeir et al., 2020).

Abbildung 3: Jährlicher Stromverbrauch von Bitcoin in annullierten TWh: In Gelb der geschätzte Wert; In Grau und hell gelb die höchste bzw. niedrigste Schätzung (Cambridge Centre for Alternative Finance, 2020)

Der Hauptgrund für den enormen Energiebedarf ist der Proof-of-Work-Mechanismus der Blockchain-Technologie, dieser basiert weder auf ineffizienten Algorithmen noch auf veralteter Hardware. Im Gegenteil, diese Blockchain-Verfahren sind zum Schutz vor Angriffen bewusst energieintensiv designt. Daher könnte auch eine deutliche Steigerung der Energieeffizienz von Mining-Hardware den Energiebedarf einer PoW-Blockchain langfristig nicht substantiell reduzieren (Sedlmeir et al., 2020). Um dieses Problem zu lösen, werden mehrere alternative Konsensmechanismen vorgeschlagen, wie zum Beispiel Proof-of-Stake (PoS) und Proof-of-Authority (PoA). Der erste ist auf dem Zufallsmechanismus aufgebaut, der bestimmt, wer den nächsten Block bauen und anhängen darf. Ohne die Entschlüsselung sinkt der ungefähre Energiebedarf pro Transaktion von 10^9 J auf 10^3 J (Sedlmeir et al., 2020). Der zweite wird als 'permissioned' Blockchain bezeichnet, die über einen Registrierungsprozess zur Identifizierung verfügt. Sie ist in der Industrie beliebt, da der ungefähre Energiebedarf pro Transaktion innerhalb eines Kleinunternehmens 1 J beträgt (Sedlmeir et al., 2020). Obwohl ein zentralisiertes System nur 0,1 J pro Transaktion verbrauchen würde, ist eine Größenordnung von 1 J pro Transaktion vernachlässigbar.

1.4 Anwendungsbeispiel der Zukunftstechnologien: Autonomes Fahren

Dieser Abschnitt beleuchtet eine der energieintensivsten Anwendungen, das autonome Fahren, das mehrere Energieverbrauchende Technologien wie Perzeptron, V2X (Connected Car), KI und weitere umfasst. Wir folgen dem Arbeitsablauf des autonomen Fahrens, um die energieintensiven Schritte zu unterscheiden:

Datenerhebung und Datenausführung (Perzeptron)

Die Ausführungen der einzelnen Komponenten eines autonomen Fahrzeugs sind zunächst mit denen eines normalen Elektrofahrzeugs vergleichbar. Der zusätzliche Energiebedarf entsteht in der

Datenerhebung zunächst über die Sensoren, welche recht energieaufwendig sind. So benötigt beispielsweise ein fortschrittlicher LiDAR-Sensor 60 W (Liu et al., 2019). Auch wenn nur einfachere Aspekte des Fahrens automatisiert werden, werden mindestens fünf einfache LiDAR-Sensoren benötigt; diese verbrauchen 12,1 W pro Sensor (Liu et al., 2019).

Datenübertragung (V2V, V2X)

Die Datenübertragung beim autonomen Fahren besteht aus zwei Teilen: der Sensordatenübertragung zum Edge-Rechner und der Kommunikation zwischen den Fahrzeugen (V2V) und allen weiteren Aspekten (V2X). Jedes Fahrzeug produziert Sensordaten zwischen 1,4 TB und 19 TB pro Stunde (Heinrich, 2017). Obwohl Fahrzeughersteller zur Senkung des Energiebedarfs Kabel, insbesondere Glasfaserkabel, für die Kommunikation im Auto verwenden, benötigt die Datenübertragung immer noch massiv Strom.

Um ein autonomes Fahren zu erreichen, ist die Kommunikation zwischen den Fahrzeugen notwendig, was zu zahlreichen Datenströmen zwischen Servern und Endgeräten führt, um den Rechenstatus in Echtzeit zu synchronisieren und die Ressourcen während der gesamten Fahrt zu verteilen. Bei einer Aufrüstung auf V2X (Vehicles to everything) werden die Datenströme in die Höhe schnellen und länger dauern. Unter Berücksichtigung der Echtzeit-Anforderung können V2V und V2X nur die 5G-Netzwerktechnologie anwenden, die nicht nur den Energiebedarf der Datenübertragung erhöht, sondern auch die Basisstationen belastet.

Datenverarbeitung und Datenspeicherung

Autonomes Fahren hat mehrere KI-Komponenten zur Entscheidung, wie zum Beispiel Bilderkennung, Videotracking, Routenplanung, Bewegungssteuerung und Spracherkennung. Daher hat jede Komponente einen ähnlichen Energiebedarf wie der oben beschriebene KI-Abschnitt. Im Vergleich zu anderen Anwendungskontexten erfordert das autonome Fahren viel höhere Genauigkeiten für jedes Modell, was größere und tiefere neuronale Netze sowie längere Versuchszeiten erfordert. Darüber hinaus steigen die Energiebedarfe durch die Echtzeitanforderung. Der eingebettete Edge-Computing-Chip muss neue Ergebnisse aus dem trainierten Modell mit aktualisierten Informationen berechnen, was für das einfache autonome Fahren mindestens 500 W verbrauchen würde (Liu et al., 2019). Für das intelligente Fahren würde diese Zahl auf 1200 W ansteigen (Liu et al., 2019).

1.5 Anwendungsbeispiel der Zukunftstechnologien: Digitaler Zwilling

Dieser Abschnitt beleuchtet eine der energieintensivsten Anwendungen, den digitalen Zwilling. Dieser ist eine digitale Repräsentanz eines materiellen oder immateriellen Objekts oder Prozesses aus der realen Welt. Es gibt mehrere Energieverbrauchende Technologien wie IoT, VR und KI mit räumlichen Netzgrafiken, um digitale Simulationsmodelle zu erstellen. Wir folgen dem Arbeitsablauf des digitalen Zwillings, um die energieintensiven Schritte zu unterscheiden:

Datenerhebung und Datenausführung (IoT)

Obwohl mit der Weiterentwicklung der Sensortechnologie die einzelne Datenerfassung immer weniger Energie kostet, ist die Datenerfassung für einen digitalen Zwilling immer noch energieintensiv (Khajavi et al., 2019). Um Stabilität zu gewährleisten, müssen redundante Sensoren eingesetzt werden. So setzen Khajavi et al. (2019) beispielsweise sieben Sensoren ein, um einen einzelnen Arbeitsplatz zu erfassen und damit die Datenzuverlässigkeit durch Sensorredundanz zu gewährleisten.

Datenübertragung (5G)

Ähnlich wie V2V und V2X beim autonomen Fahren, werden auch hier große Datenmengen übertragen. Obwohl jeder Sensor normalerweise nur Daten in der Größenordnung von Bytes sendet, halten zahlreiche Endgeräte die Datenübertragung zwischen Servern und sich selbst während ihrer gesamten Lebensdauer aufrecht. Dies erhöht die massiven Datenströme, um den Rechenstatus in Echtzeit zu synchronisieren und die Ressourcen zu verteilen. Darüber hinaus muss VR als wichtiger Weg zur Visualisierung von Simulationsergebnissen digitaler Zwillinge Modelle in der Größenordnung von MB pro Szenario übertragen, was ebenfalls eine große Herausforderung für den Energiebedarf darstellt, insbesondere auf der Basis von 5G-Mobilfunknetzen.

Datenverarbeitung und Datenspeicherung (BIM, KI, VR)

Als Basis des digitalen Zwillings nimmt das auf Sensordaten basierende 3D-Gebäudemodell einen großen Teil des Energiebedarfs ein, was nicht nur an den massiven Daten, sondern auch an den komplexen Zusammenhängen liegt. Auf der Grundlage der Gebäudemodellierung erhöht das VR-Rendering, wie im vorigen Abschnitt erwähnt, die Berechnungszeiten dramatisch. Glücklicherweise sind angewandte KI-Methoden zur intelligenten Diagnose verhältnismäßig einfach und benötigen daher selten einen langen Trainingsprozess, während 'deep learning' Algorithmen deutlich energieintensiver sind.

2 Nachhaltigere Lösungsansätze

Im Hinblick auf die oben genannten drei Schritte geben wir hier einen Überblick über verfügbare allgemeine Lösungen aus Hardware (Datenerfassung und -ausführung), Protokollen (Datenübertragung) und Algorithmen (Berechnung).

2.1 Hardware-Lösungen

Gegenwärtig sind effiziente Lösungen im Hardwarebereich bereits verbreitet, insbesondere für IoT-Sensoren und Trainings für maschinelles Lernen, die das BMBF ebenfalls schwerpunktmäßig fördern wird. Im Hinblick auf die KI-Strategie der EU sollte die zukünftige KI auf den industriellen Stärken Europas aufbauen. Hardware-Lösungen wären nach wie vor ein Forschungsschwerpunkt (Hernandez & Brown, 2020).

Generell gibt es für Energieeinsparungen bei Sensoren folgende Möglichkeiten:

- Minimierung der Verarbeitungstätigkeiten von Sensoren: Deaktivierung einiger nutzloser Funktionen für aktuelle Anwendungen, Zuweisung spezifischer Aufgaben an bestimmte Knoten/Schichten/Kerne durch optimierte Zeitplanung und Umstellung auf passiven RFID-Modus anstelle des aktiven Modus (Arshad et al., 2017)
- Ersetzen von Sensoren durch Chips (Arshad et al., 2017)
- Anwendung von Cache zur Speicherung früherer Informationen oder benachbarter Informationen zur Reduzierung der Datenübertragungsmenge (Xu et al., 2020)

Was die energieeffiziente KI-Hardware betrifft, so konzentrierte sie sich ursprünglich auf die Vergrößerung des In-Chip-Speichers und einen effizienteren Speichertransfer. Hier hatte beispielsweise die Parallelisierung des Datentransfers vom In-Chip-Speicher auf Disks, die vorübergehende Speicherung von Pixeln auf dem Chip anstatt des Weitertransfers (Wang et al., 2017) sowie die Maximierung der Wiederverwendung von Eingabemerkmalen (Capra et al., 2020) Priorität. In letzter Zeit wendet sich die Forschung vermehrt der Beschleunigung von Prozessen mit geringer Ausnutzung und mit variabler Bitbreite zu. Während der erstgenannte Bereich die hohe Anzahl von Nullen in den Matrizen ausnutzt, nutzt der zweite die komprimierte Darstellung von Zahlen, z. B. von 32-Bit-Zahlen in Richtung 16-Bit- oder 8-Bit-Zahlen (Capra et al., 2020).

2.2 Protokoll Lösungen

Dieser Abschnitt befasst sich mit den verfügbaren nachhaltigeren Kommunikationsprotokollen, insbesondere mit der Frage, wie 5G weniger Energie verbrauchen kann. Folgende Punkte eröffnen einen Lösungsraum:

- Nutzung des ungenutzten und unlizenzierten Spektrums: *milli-meter-wave communication* (mmWave) und *Long Term Evolution* (LTE) im unlizenzierten Spektrum (LTE-U)
- Verringerung des Abstandes zwischen Sender und Empfänger (Tx-Rx) und Verbesserung der Wiederverwendung von Frequenzen: *ultra-dense networks* UDNs und Geräte-zu-Geräte-Kommunikation (*device-to-device communication* D2D)
- Verbesserung der spektralen Effizienz (SE) durch den Einsatz einer großen Anzahl von Antennen: *massive multiple-input multiple-output* (M-MIMO) (Wu et al., 2017)

Abgesehen von den 5G-Kommunikationsprotokollen, wie in Abschnitt 1.3 erwähnt, konzentriert sich die Blockchain-Technologie auch auf ihre Konsensprotokollerweiterung ('consensus protocol enhancement'). Obwohl 'Proof-of-Authority' einen akzeptablen Energiebedarf hat, ist er auf bestimmte Netzwerkgrößen beschränkt. Daher gibt es mehr Konsensmechanismen, die auf vertrauenswürdigen Zufallszahlengeneratoren durch sichere Hardware-Module basieren, wie zum Beispiel Proof-of-lapsed-time, die das Skalierbarkeits-Trilemma lösen könnten, um den besten Kompromiss zwischen Leistung, Sicherheit und Energiebedarf zu finden (Sedlmeir et al., 2020).

2.3 Algorithmische Lösungen

Nachhaltige Algorithmuslösungen konzentrieren sich insbesondere auf 1) maximale Lerngenauigkeit bei minimalem Rechenaufwand und 2) effiziente Verarbeitung großer Datenmengen (Al-Jarrah et al., 2015). Um die beiden oben genannten Ziele zu erreichen, lassen sich die Lösungen für KI-Methoden aufteilen in Dateneffizienz (weniger Iterationen zum Trainieren erforderlich) und Reduzierung der Anzahl der pro Iteration erforderlichen Berechnungen (Hernandez & Brown, 2020). Im Hinblick auf die drei in Abschnitt 1.1 analysierten Schritte entlang der Datenverarbeitungskette gibt es drei Hauptlösungen:

- Die Komplexität der Modelle: Einsatz verschiedener Techniken zur Reduktion von Lernparametern, wie Residualverbindungen, Sparsamkeit, Batch-Normalisierung und angemessene Skalierung der Architektur neuronaler Netze (Hernandez & Brown, 2020).
- Der Bedarf an riesigen Datenmengen: Die Anwendung einer lokalen Lernstrategie, die die Trainingsstichproben in Cluster aufteilt, baut damit für jedes Cluster ein eigenes lokales Modell auf,

sodass bei ähnlicher, noch besserer Performance die Modelle nicht auf einem großen Datensatz trainiert werden müssen (Al-Jarrah et al., 2015).
- Die Vielfalt der Modellauswahl: Automatisieren der Suche nach der am besten geeigneten neuralen Netzwerkstruktur, anstatt die Iteration auf der von Hand entworfenen Architektur durchzuführen (Hernandez & Brown, 2020). Mit der intelligenten Architektursuche lassen sich Trainingsexperimente reduzieren.

Darüber hinaus beginnt die Industrie, wie beispielsweise Google und Amazon, KI-Methoden zur Optimierung des Berechnungsprozesses und des Kühlplans des Rechenzentrums zu entwickeln.

3 Herausforderungen

Es werden insbesondere zwei Herausforderungen auf dem Weg zu energieeffizienten Zukunftstechnologien identifiziert:

1. Während die Effizienz von Algorithmen bereits inhärent in der Programmierung berücksichtigt wird, ist es derzeit noch schwierig den gesamten Energiebedarf von KI(-Anwendungen) standardisiert zu bestimmen, da es noch keine Standards zu optimalen Samplegrößen, optimalen Anzahl an Trainingsrunden etc. gibt. Hinzu kommt, dass die Nettobetrachtung nur sektorenübergreifend durchgeführt werden kann. Auch hierzu fehlt derzeit noch eine standardisierte Methodik.
2. Die weltweit führenden Algorithmen gehen mit riesigen und tiefen neuronalen Netzen einher, deren Training und Betrieb sehr viel Energie kostet. Deutschland liegt derzeit nicht an der Spitze der KI-Entwicklung. Um zur Konkurrenz aufzuschließen, sind die deutsche Forschung und deutsche Unternehmen in erster Linie darum bemüht, an diese Leistungen anzuschließen und nicht darum, möglichst energiesparende Algorithmen zu entwickeln (siehe Fragen für die Diskussion).

4 Handlungsansätze für die Energieeffizienzpolitik

Sieben mögliche Handlungsansätze werden hier identifiziert:

1. Energieeffiziente KI sowohl in der **deutschen KI-Strategie als auch in der KI-Strategie der EU** verankern. Dem Energiebedarf der KI wird derzeit in der deutschen KI-Strategie noch keine (erhöhte) Beachtung geschenkt (Bundesregierung, 2018)
2. Als **Steuerungsinstrument** sollte ein über Technologien und Infrastruktur hinwegreichendes **Lagebild** kontinuierlich erhoben und weiterentwickelt werden, so dass Handlungsempfehlungen abgeleitet und Entwicklungen justiert werden können.
3. **Studie zum Nettonutzen** von digitalen Technologien zur wissenschaftlichen Aufbereitung der Thematik. Die Studie könnte u.a. Grundlagen für die Schaffung eines Standards zur Messung der Energieeffizienz von Algorithmen inklusive einer standardisierten Methode zur Nettobetrachtung sektorübergreifender Einsparungen durch digitale Technologien legen.
4. **Aufmerksamkeit in der akademischen Welt** erhöhen, um Forschung über energieeffiziente innovative digitale Technologien und/oder intelligente Energieplanungssysteme für zukünftige digitale Technologien zu ermöglichen.

Dies kann beispielsweise durch gezielte Forschungsförderung und Pilotprojekte wie des Pilotinnovationswettbewerbs "Energieeffizientes KI-System" des BMBF geschehen. "Das BMBF fördert […] Hochschulen und öffentliche Forschungseinrichtungen, damit sie ihre Ideen für energieeffiziente Elektronik-Hardware für Künstliche Intelligenz (KI) in einem Testaufbau zeigen und sich damit in einer Leistungsmessung vergleichen. […] Im Erfolgsfall wird damit eine Sprunginnovation für KI-Elektronik-Anwendungen ausgelöst."[1]

5. **Sensibilisierung in der Industrie** für den erhöhten Energiebedarf durch innovative digitale Technologien und Wissensaufbau zu möglichen Energieeffizienzpotenzialen.
6. Zusammenbringen der **KI-Community mit dem Energiesektor**/mit Energieexpert*innen (Verknüpfung von digitalen Technologien mit mehr erneuerbaren Energien und energieeffizienten Dienstleistungen).
   Ein Beispiel für diesen Ansatz ist das durch das BMWi geförderte Pilotierungs- und Vernetzungslabor "Future Energy Lab" der dena[2]*, welches zum Ziel hat Unternehmen zu vernetzen um den Einsatz von innovativen digitalen Technologien im Energiesektor voranzutreiben.*
7. **KI-Unternehmen**, insbesondere Start-ups, für den Energiesektor gewinnen und fit machen

5 Leitfragen für die Diskussion

- *Teilen Sie die **Darstellungen in diesem Inputpapier** zum Energiebedarf einzelner Technologien/Anwendungen?*
- *Welche digitalen Technologien und deren Anwendungsfälle sollte das BMWi im Rahmen einer **Studie zur Energieeffizienz von Zukunftstechnologien** in den Blick nehmen?*
- *In welchen Bereichen sehen Sie konkreten Forschungsbedarf?*
- *Kann eine **Nettobetrachtung** (neue Verbräuche minus Einsparungen an anderer Stelle) einzelner Technologien und Anwendungsfällen sinnvoll sein, oder sind die daraus resultierenden Bandbreiten zu groß?*
- *Sind Ihnen **Ansätze zur energieeffizienten Ausgestaltung digitaler Zukunftstechnologien** bekannt, die das BMWi pilotieren sollte? Sind Ihnen Ansätze zur energieeffizienten Ausgestaltung digitaler Zukunftstechnologien aus anderen Ländern bekannt, die auch Relevanz für Deutschland/Europa haben können (Forschung, Start-ups, Unternehmen, etc.)? Gibt es Ansätze von denen Sie denken, dass Sie in Deutschland/Europa nicht verfolgt werden sollten – wenn ja warum?*
- *Könnte eine energieeffiziente KI ein **Qualitätsmerkmal für KI Made in Germany** werden? Was bedeutet für Sie energieeffiziente KI, welche Kriterien muss eine KI-Anwendung erfüllen um energieeffizient zu sein?*
- *Kann eine **Einordnung** von KI-Anwendungen hinsichtlich ihrer **Kritikalität** sinnvoll sein, um zu analysieren, für welche Anwendungen ein erhöhter Energiebedarf gerechtfertigt sein kann?*
- *Welche Argumente sprechen für und welche **gegen Ökodesign- und Energiebedarfskennzeichnungs-Ansätze** für Hardware- und Softwarelösungen sowie digitale Dienstleistungen?*

[1] https://www.bmbf.de/foerderungen/bekanntmachung-2371.html
[2] https://www.dena.de/future-energy-lab/

6 Literatur

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87-93.

Arshad, R., Zahoor, S., Shah, M. A., Wahid, A., & Yu, H. (2017). Green IoT: An investigation on energy saving practices for 2020 and beyond. *IEEE Access*, *5*, 15667-15681.

Bitkom. 2020. Klimaschutz durch digitale Technologien – Chancen und Risiken. Kurzstudie.

Bundesregierung. 2018. Strategie Künstliche Intelligenz der Bundesregierung.

Cambridge Centre for Alternative Finance (2020). Cambridge Bitcoin Electricity Consumption Index. www.cbeci.org Zuletzt abgerufen am 27.9.2020

Capra, M., Bussolino, B., Marchisio, A., Shafique, M., Masera, G., & Martina, M. (2020). An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks. *Future Internet*, *12*(7), 113.

Fadeyi, O., Krejcar, Maresova, P., Kuca, K., Brida, P., & Selamat, A. (2019). Opinions on Sustainability of Smart Cities in the Context of Energy Challenges Posed by Cryptocurrency Mining. *Sustainability 12*.

Heinrich, S. (2017). Flash Memory in the emerging age of autonomy. In *Flash Memory Summit 2017*.

Hernandez, D., & Brown, T. B. (2020). Measuring the Algorithmic Efficiency of Neural Networks. *arXiv preprint arXiv:2005.04305*.

Khajavi, S. H., Motlagh, N. H., Jaribion, A., Werner, L. C., & Holmström, J. (2019). Digital twin: vision, benefits, boundaries, and creation for buildings. *IEEE Access*, *7*, 147406-147419.

Leng, Y., Chen, C. C., Sun, Q., Huang, J., & Zhu, Y. (2019, June). Energy-efficient video processing for virtual reality. In *Proceedings of the 46th International Symposium on Computer Architecture* (pp. 91-103).

Liu, Z., Tan, H., Kuang, X., Hao, H., & Zhao, F. (2019). The Negative Impact of Vehicular Intelligence on Energy Consumption. *Journal of Advanced Transportation*, *2019*.

OpenAI. 2018. AI and Compute. Retrieved from: https://openai.com/blog/ai-and-compute/ [accessed on September 21st 2020]

Sedlmeir, J., Buhl, H. U., Fridgen, G., & Keller, R. (2020). The energy consumption of blockchain technology: beyond myth. *Business & Information Systems Engineering*, 1-10.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

Wang, J., Lin, J., & Wang, Z. (2017). Efficient hardware architectures for deep convolutional neural network. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *65*(6), 1941-1953.

Wang, Y., Wang, Q., Shi, S., He, X., Tang, Z., Zhao, K., & Chu, X. (2020, May). Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)* (pp. 744-751). IEEE.

Wu, Q., Li, G. Y., Chen, W., Ng, D. W. K., & Schober, R. (2017). An overview of sustainable green 5G networks. *IEEE Wireless Communications*, *24*(4), 72-80.

Xu, C., Wang, X., Yang, H. H., Sun, H., & Quek, T. Q. (2020). AoI and Energy Consumption Oriented Dynamic Status Updating in Caching Enabled IoT Networks. *arXiv preprint arXiv:2003.00383*.

7 Team

Jiao Jiao – Fraunhofer ISI – jiao.jiao@isi.fraunhofer.de

Dr. Heike Brugger – Fraunhofer ISI – heike.brugger@isi.fraunhofer.de

i

Es liegt in der Natur des Themengebietes Zukunftstechnologien, dass die verfügbare Datenlage dünn und vorhandene Studien noch sehr beschränkt verfügbar und mit Unsicherheiten behaftet sind. Dieses Papier bietet einen Überblick über die zentralen zu betrachtenden Aspekte basierend auf dem derzeitigen Stand der Erkenntnis, ohne den Anspruch das insbesondere die abgeschätzten Energiebedarfe bereits abschließend quantifiziert sind.