EDGE AI: Das verfügbare Potenzial ist grenzenlos ein Interview mit Wolfgang Ecker von Infineon

2024/07/08

Aus Aus Industry.zero & Transformation for Industy Leaders am 9.7.2024 [1] am 9.7.2024

Die jüngsten Erfolge in der generativen KI basieren auf einem Anstieg zentral verarbeiteter Daten, größeren neuronalen Netzen und mehr Rechenkapazität. Dies wirft Fragen zu Datenschutz, Kosten und Ressourcenverbrauch auf. Daher wird parallel ein anderer Ansatz verfolgt: die Dezentralisierung von KI-Architekturen nach dem Vorbild des Edge Computing – genannt Edge AI. Das Ziel: Daten nahe am Nutzenden und nicht in der Cloud verarbeiten. Wolfgang Ecker, Distinguished Engineer bei Infineon Technologies und Honorarprofessor der TU München, erklärt Vorteile, Einsatzmöglichkeiten und aktuellen Hürden von Edge AI.

"Herr Ecker, wo liegen die größten Vorteile von Edge AI gegenüber klassischen Cloud-Ansätzen?"

Wolfgang Ecker: "Lassen Sie es mich einmal so sagen: Aus technologischer Sicht ist Edge AI im Vergleich zur Cloud erst einmal eine zusätzliche technische Herausforderung. Die Berechnung der Netze muss bei Edge AI mit Milli-Watt elektrischer Leistung auskommen, bei der Cloud werden Kilo- oder Megawatt verbraucht. Auf der Kostenseite ist man bei Edge AI eher im Euro-Bereich, in der Cloud bei Tausenden und Millionen von Euro. Entsprechend müssen die KI-Recheneinheiten kleiner sein und mit weniger Strom auskommen, was nur durch besonders optimierte Netze möglich ist. Deshalb liegen die technischen Vorteile der Edge-AI-Lösungen in der Anwendung der Technologie. Edge AI muss die Daten nicht erst an die Cloud schicken und auf eine Antwort warten, sondern kann nahe am Auftreten der Daten ausgeführt werden. Schnellere und garantierte Antworten der KI sind deshalb ebenso ein technischer Vorteil wie der Schutz der Daten, da diese nur lokal vorgehalten werden müssen. Die Anwendungen sind auch robuster, da ein Ausfall der Kommunikation mit der Cloud nicht in Betracht gezogen werden muss. Zuletzt haben Edge-AI-Anwendungen einen viel geringeren CO2-Abdruck als Anwendungen in der Cloud."

"Welche Potenziale ergeben sich daraus – und wo stehen wir in Deutschland in puncto Transfer?"

Wolfgang Ecker: "Die genannten Vorteile von kleinem Formfaktor, geringen Kosten, geringem Energieverbrauch, besser geschützten Daten und systemisch inhärent robusteren – da unabhängigeren – Implementierungen öffnen eine Vielzahl von Potentialen gerade in deutschen Leitindustrien wie Automobil, Maschinenbau und Medizintechnik. Ein Beispiel ist die Vehicle-to-Vehicle-Kommunikation beim teilautonomen Fahren: Durch Sensordaten des Autos (zum Beispiel Lidar, Kameras, Radar) sowie Verkehrsdaten, die über Kommunikationsnetze zwischen Fahrzeugen ausgetauscht werden, können lokale KI-Modelle zum Einsatz kommen, die diese eingehenden Daten zuverlässig und in Echtzeit verarbeiten und so Anomalien erkennen. Bei gefährlichen Situationen können so Warnungen kommuniziert oder gar Maßnahmen unabhängig eingeleitet werden, um einen Unfall zu vermeiden. Ein weiteres Anwendungsfeld ist die Industrierobotik: Edge AI kann hier mithilfe des föderierten Lernens so umgesetzt werden, dass Kommissionier-Roboter in der Lage sind, mit KI zu „fühlen“ und voneinander zu lernen und so auch unbekannte Objekte zuverlässig zu greifen. Meiner Meinung nach sind die Opportunitäten der Edge AI grenzenlos. Auch wenn es bereits Erfolge vorzuweisen gibt, so nutzen wir das verfügbare Potenzial aber bei weitem noch nicht aus. Lokale Ansätze und Gräben um eigene Arbeitsgebiete verhindern eine holistische Herangehensweise. Die Gestaltung der Netze, das Trainieren der Netze, die Übersetzung der Netze und die Hardware-Architekturen zur Berechnung der Netze werden weitgehend unabhängig betrachtet. Oft werden Lösungen aus der Cloud AI angepasst statt passgenaue Edge-AI-Komplettlösungen zu erarbeiten. Eine holistische Herangehensweise aber ist notwendig, um eine leistungsstarke Edge-AI-Technik bereitzustellen. Und ebenso müssen Technik und Anwendungen gemeinsam betrachtet werden. Nur mit dem Wissen der Anwendung können die Edge-AI-Maschinen effizient gestaltet werden und im Gegenzug können nur mit dem Wissen der Leistungsfähigkeit der Edge-AI-Technik neue Anwendungen entwickelt werden."

"Werden sich in absehbarer Zeit auch große Sprachmodelle „on the edge“ ausführen und/oder trainieren lassen? Was ist dafür nötig?"

Wolfgang Ecker: "Wörtlich genommen denke ich, wird das nie klappen. „Große Modelle“ auf der einen Seite sowie Energie- sowie Kosten-Effizienz auf der anderen Seite passen nicht zusammen. Entscheidend wird sein, ob es klappt, die großen Sprachmodelle so zu skalieren, dass sie von Edge Devices verarbeitet werden können. Ebenso ist es wichtig, das Training so effizient zu gestalten, dass es in der Edge ausgeführt werden kann. Erste Ansätze sind bekannt, also warum soll es mit der oben dargestellten holistischen Herangehensweise nicht klappen? Es muss aber klar gemacht werden, dass eine Edge-AI-Umsetzung nie die Universalität und Leistungsfähigkeit einer Cloud AI erreichen werden. Entsprechend bergen verteilte und/oder gemischte Edge-AI- / Cloud-AI-Lösungen ein weiteres großes Potential. Vielleicht wandern neue Edge-AI-Techniken auch zurück in die Cloud, um dort den Energiebedarf und den CO2-Abdruck zu reduzieren."

The Scale4Edge project (project label 16ME0122K-140, 16ME0465, 16ME0900, 16ME0901) is supported by the German Federal Ministry of Research, Technology and Space (BMFTR).

Links:
[1] https://www.industr.com/de/edge-ai-das-verfuegbare-potenzial-ist-grenzenlos-2759326