Fraunhofer
IAIS

FRAUNHOFER-INSTITUT FÜR INTELLIGENTE ANALYSE- UND INFORMATIONSSYSTEME IAIS

MASCHINELLES LERNEN »ON THE EDGE«

Konzepte und Vorteile am Beispiel des autonomen Fahrens

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Schloss Birlinghoven
53757 Sankt Augustin

Ansprechpartner

Dr. Tim Wirtz
Telefon +49 2241 14-3511
tim.wirtz@iais.fraunhofer.de

www.iais.fraunhofer.de

Wie unterscheidet sich maschinelles Lernen »on the edge« vom Lernen in der Cloud?

Hochautomatisierte Fahrzeuge erkennen Verkehrsschilder, halten Abstand zu anderen Fahrzeugen, bremsen vor Hindernissen rechtzeitig und finden ihren Weg zum Ziel ohne viel menschliches Zutun. Die technologische Entwicklung in der Künstlichen Intelligenz (KI) sorgt dafür, dass dieses Zukunftsszenario in greifbare Nähe rückt. Insbesondere mit Deep-Learning – einer Form des maschinellen Lernens auf großen Datenmengen, die mit tiefen neuronalen Netzen als Modellen arbeitet – erzielt die Forschung zurzeit vielversprechende Fortschritte.

Bisher findet das Training der neuronalen Netze vorwiegend in der Cloud statt: Auf einer zentralen Big-Data-Plattform werden historische Daten kontinuierlich durch neue Datensätze ergänzt, die im Einsatz gewonnen werden. Das Netz wird auf diesen Daten trainiert und verbleibt auch im Einsatz in der Cloud, so dass Anfragen an das trainierte Modell ebenfalls dorthin übermittelt werden.

Dieses Paradigma ist jedoch in zahlreichen Anwendungen weder technisch wünschenswert noch rechtlich möglich: Wie im Gesundheitswesen spielt auch im Automotive-Umfeld der Datenschutz eine besondere Rolle und stellt einen gewichtigen Grund dar, um Kundendaten nicht an eine zentrale Plattform in der Cloud zu übermitteln und dort zu speichern. Hinzu kommen technische Hindernisse wie geringe Bandbreiten, hohe Kommunikationskosten, Gefahren des Datenabgriffs durch Cyberattacken sowie die Notwendigkeit kurzer Reaktionszeiten.

Ein Ansatz, um diesen Anforderungen gerecht zu werden, ist das verteilte Lernen. Dabei werden keine Kundendaten in die Cloud übertragen und das Training der Netze findet auf den Endgeräten statt. Im Falle des autonomen Fahrens entsprechen die Endgeräte den Fahrzeugen einer Flotte, weshalb auch von Flottenlernen[1] (engl. »fleet learning«) gesprochen wird.

Forscherinnen und Forscher des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS und des Volkswagen-Konzerns haben gemeinsam untersucht, wie die zentralen Herausforderungen des Flottenlernens bewältigt werden können. Auf der European Conference on Machine Learning (ECML) 2018 in Dublin haben sie ihre Ergebnisse vorgestellt und das Paper »Efficient Decentralized Deep Learning by Dynamic Model Averaging« veröffentlicht.

Wie verteiltes Lernen funktioniert, welche Herausforderungen dabei bestehen und welche Fortschritte bereits erzielt wurden, ist hier in Kürze zusammengefasst.

Downloads:

Efficient Decentralized Deep Learning by Dynamic Model Averaging:
www.ecmlpkdd2018.org/wp-content/uploads/2018/09/182.pdf

Introducing Noise in Decentralized Training of Neural Networks:
dmle.iais.fraunhofer.de/Papers.html

Was sind die Ziele des verteilten Lernens »on the edge«?

Der Erfolg des maschinellen Lernens hängt oft von der Verfügbarkeit vieler Trainingsdaten (Big Data) ab – jedoch ist der Speicherplatz auf einem einzelnen Endgerät gering. Verteiltes Lernen erfordert folglich Konzepte, um Deep-Learning-Modelle auf den datenerzeugenden Endgeräten oder Fahrzeugen zu trainieren und anzuwenden, also an der »Edge« statt in der »Cloud«.

Die zentrale wissenschaftliche Fragestellung dabei lautet: Wie können die Modelle auf einem lokalen Gerät von den Zusammenhängen profitieren, die auf anderen Geräten gelernt werden, ohne dass Rohdaten ausgetauscht oder in der Cloud zentralisiert werden müssen?

Ein Beispiel aus der Spracherkennung veranschaulicht diese Problematik: Auf einem einzelnen Smartphone wird die Spracherkennung für dessen Nutzerinnen und Nutzer optimiert, also für eine junge oder alte Stimme, einen Dialekt oder den Akzent eines Nicht-Muttersprachlers. Um von den lokalen Modellen für die jeweiligen Gerätebenutzer zu einem einzigen Modell zu gelangen, das eine Vielzahl an Stimmvarianten erfolgreich verarbeiten kann, sollen die gelernten Erkenntnisse zwischen den Modellen ausgetauscht werden. Am Ende beherrscht ein einziges gemeinsames Modell auf allen Endgeräten die Verarbeitung vieler Stimmtypen. Doch wie realisiert man diesen Erkenntnisaustausch, ohne dafür auch nur eine einzige Stimmaufzeichnung in die Cloud zu übertragen?

Übertragen auf den Automobil-Kontext bedeutet dies, ein stabiles gemeinsames Modell zu trainieren, welches zum Beispiel vorausschauende Wartung für Fahrzeuge ermöglicht, die unterschiedlich beansprucht und gepflegt werden und in unterschiedlichen Gegenden und Wetterverhältnissen betrieben werden. Ein weiterer Anwendungsfall ist die Vorhersage von Fahreraktionen, das zentrale Element des automatisierten Fahrens: Dabei lernt ein tiefes neuronales Netz beispielsweise, aus den Bildern von Bordkameras die optimalen Reaktionen vorherzusagen – etwa Lenkradeinschlag, Beschleunigen, Bremsen oder Spurwechsel. Die zugrundeliegenden Daten wie Kamerabilder oder 3D-Laser-Scans sollen das Fahrzeug dabei nicht verlassen.

Techniken und Verfahren für dieses Lernen an der Edge werden bereits seit einigen Jahren entwickelt, allerdings unter wechselnden Bezeichnungen wie ressourcenbeschränktes, verteiltes, dezentrales, kollaboratives oder kommunikationseffizientes Lernen.

Viele Grundlagen dafür wurden in den vom Fraunhofer IAIS geleiteten EU-Forschungsprojekten »Local InFerence in massively distribuTed Systems« (LIFT) und »Flexible Event pRocessing for big dAta aRchItectures« (FERARI) erarbeitet.[2]

Wie funktioniert verteiltes Lernen?

Während beim zentralisierten Lernen – dem Lernen in der Cloud – der Lernalgorithmus sein Modell auf der zentralen Plattform mit den dort angehäuften Daten trainiert,

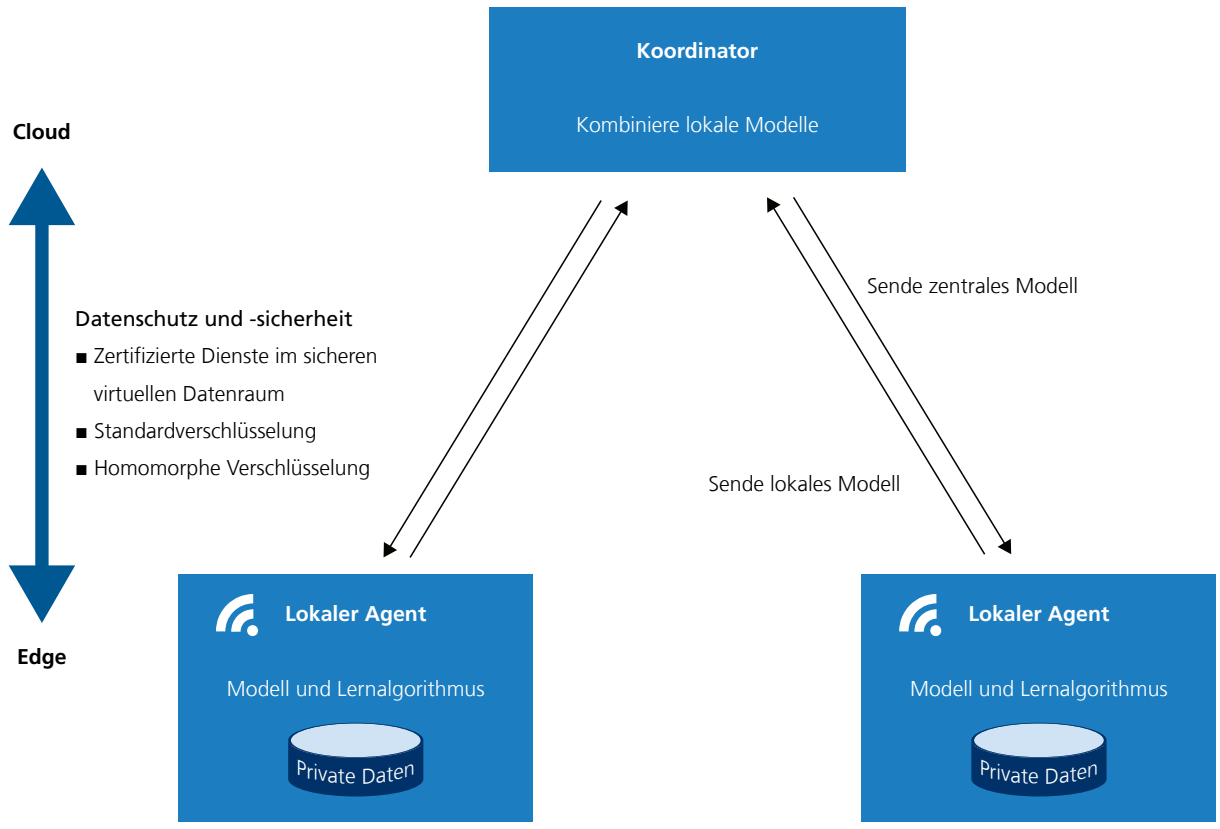
liegen Algorithmus und Modell beim verteilten Lernen – dem Lernen an der Edge – lokal auf oder nahe bei jedem Gerät. Um von den Daten aller Geräte zu profitieren, werden die lokal gelernten Modelle gesammelt, zu einem gemeinsamen Modell zusammengeführt und dieses synchronisierte Modell wieder an alle Geräte verteilt (vgl. Bild 1).

Die zentrale Idee dieses Vorgehens ist es, trainierte Modelle anstelle von Rohdaten auszutauschen. Die Daten werden ausschließlich für lokales Modell-Training verwendet. Der Aufwand für den Austausch von Modellen ist im Vergleich zu den Rohdaten deutlich geringer. Neue Rohdaten werden vor Ort ohne Verzögerung verarbeitet und lange Übertragungszeiten vermieden.

Das Prinzip des Lernens an der Edge lässt offen, ob lokale Modelle in regelmäßigen Abständen oder dynamisch nach Bedarf ausgetauscht werden. Der technisch anspruchsvollere dynamische Ansatz zielt darauf ab, unnötigen Austausch von Modellen zu vermeiden und synchronisiert sie nur dann, wenn tatsächlich eine Verbesserung erwartet wird.

Für traditionelle lineare Modelle, wie zum Beispiel Regressionsmodelle, ist bereits bekannt, dass dieser dynamische Austausch erfolgreiches Modelltraining ermöglicht.[2] Aber auch komplexere kernel-basierte Verfahren und Markov-Modelle zur Vorhersage von komplexen Ereignissen[3] können dynamisch synchronisiert werden, wie Forscherinnen und Forscher des Fraunhofer IAIS gezeigt haben.

Bild 1: Kommunikation zwischen Koordinator und lokalen Agenten



Welche Fortschritte gibt es bei verteiltem Deep Learning?

Ein Forscherteam von Google hat im Jahr 2016 erstmals tiefe neuronale Netze unter Verwendung regelmäßiger Synchronisation erfolgreich verteilt trainiert.[4] Aufbauend auf diesen Ergebnissen hat das Fraunhofer-Team zusammen mit Forschern des Volkswagen-Konzerns untersucht, wie gut sich tiefe neuronale Netze mittels dynamischer Synchronisation verteilt trainieren lassen.[1]

Zur Bewertung der verteilten Trainingsmethoden haben die KI-Experten zwei Kenngrößen betrachtet: den notwendigen Aufwand für den Austausch der Modelle und die erreichte Qualität des final trainierten Gesamtmodells. Der Erfolg: Mit dynamischer Synchronisation wurde der Kommunikationsaufwand deutlich reduziert, während die Modelle nur marginal an Qualität eingebüßt haben (vgl. Bild 2). Die Qualität ließ sich durch zusätzliche Trainingsiterationen jedoch wieder steigern.

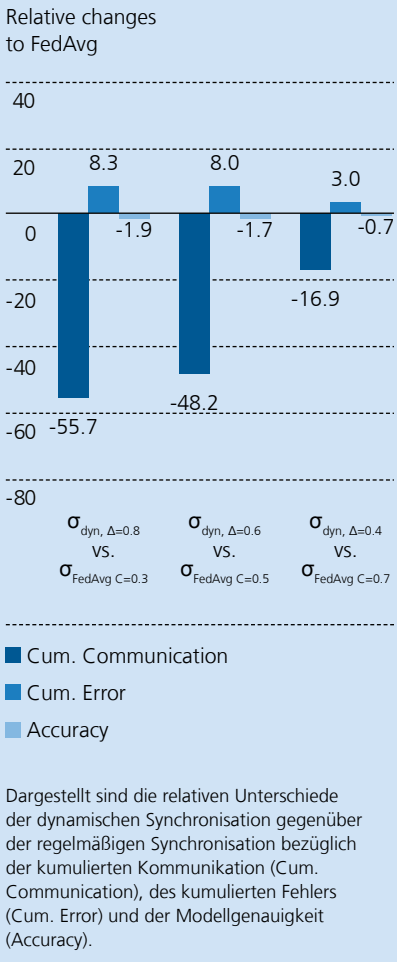
Weitere Qualitätssteigerungen erzielten die Teams von Volkswagen und Fraunhofer IAIS, indem sie während des verteilten Trainingsprozesses sogenanntes »Rauschen« hinzufügten. Das sind kleine, zufällige Modifikationen der Modellparameter.[5] Im zentralisierten Lernen ist dies eine verbreitete Methode zum Trainieren robuster Modelle, die auch auf neuen Daten gute Ergebnisse liefern. Im verteilten Lernen von tiefen neuronalen Netzen konnte dieser Effekt ebenfalls verstärkt beobachtet werden: Experimente zeigten, dass das »Verrauschen« von Netzwerkgewichten die Qualität der verteilt trainierten Modelle wesentlich verbessert.

Ausblick

Das Lernen an der Edge bietet im Vergleich zum Lernen in der Cloud wichtige Vorteile für zahlreiche Anwendungen der Künstlichen Intelligenz – nicht zuletzt durch das datenschutzfreundliche Trainingskonzept und die erheblichen Einsparungen an Kommunikationsaufwand und -kosten.

Das Lernen in autonomen Fahrzeugflotten stellt dabei nur eines von vielen Einsatzgebieten dar. Im Maschinenbau etwa ermöglicht verteiltes Lernen von Modellen, dass Hersteller smarte Services für ausgelieferte Maschinen anbieten können, ohne dafür auf die Rohdaten ihrer Kunden zugreifen zu müssen. Auch im biomedizinischen Bereich wird dieser Ansatz bereits genutzt, zum Beispiel um für verschiedene Kliniken ein gemeinsames Modell zur Tumorerkennung zu lernen. Die sensiblen Patientendaten verlassen die jeweilige Klinik nicht.[6] Die Gegebenheiten solcher Realeinsätze besser zu berücksichtigen und die Datenschutzfreundlichkeit des Ansatzes weiter zu erhöhen[7] sind Schwerpunkte der aktuellen Forschungstätigkeiten im Bereich des verteilten Lernens.

Bild 2: Vergleich von dynamischer ($\sigma_{dyn,\Delta}$) und regelmäßiger ($\sigma_{FedAvg,C}$) Synchronisation.



Dargestellt sind die relativen Unterschiede der dynamischen Synchronisation gegenüber der regelmäßigen Synchronisation bezüglich der kumulierten Kommunikation (Cum. Communication), des kumulierten Fehlers (Cum. Error) und der Modellgenauigkeit (Accuracy).

Die Autorinnen und Autoren

M.Sc. Joachim Sicking arbeitet am Fraunhofer IAIS als Data Scientist. Er forscht an Methoden des verteilten maschinellen Lernens und deren Anwendung im Bereich des autonomen Fahrens. Ein weiterer Schwerpunkt seiner Arbeit sind Machine-Learning-Projekte im Bereich Industrie 4.0 mit einem Fokus auf Fragestellungen der Ausfallprädiktion und der Produktionsoptimierung.

Dr. rer. nat. Tim Wirtz ist stellvertretender Abteilungsleiter am Fraunhofer IAIS. Dort arbeitet er sehr intensiv an den Anwendungen der künstlichen Intelligenz in Bereichen wie dem autonomen Fahren und der Industrie 4.0. In dieser Rolle verantwortet er außerdem das Thema AI-Strategy.

Dr. rer. nat. Angi Voss arbeitet in der Geschäftsstelle der Kompetenzplattform KI.NRW und beschäftigt sich dort mit der Zertifizierung von KI. In der Geschäftsstelle der Fraunhofer-Allianz Big Data ist sie verantwortlich für die Konzeption des Data-Scientist-Schulungs- und Zertifizierungsprogramms. Seit 2013 hat sie an Potenzialanalysen zu Big Data, Künstlicher Intelligenz und Maschinellem Lernen in Deutschland mitgewirkt.

M.Sc. Nathalie Paul ist Mathematikerin und arbeitet als Data Scientist in der Abteilung Knowledge Discovery am Fraunhofer IAIS. Ihr Forschungsschwerpunkt liegt auf dem verteilten Lernen mit Fokus auf Verfahren, welche die Privatsphäre bewahren und den Datenschutz gewährleisten. Sie beschäftigt sich dabei insbesondere mit Anwendungen im Bereich Industrie 4.0.

Literatur

1 Kamp, M., Adilova, L., Sicking, J., Hüger, F., Schlicht, P., Wirtz, T., Wrobel, S.: Efficient Decentralized Deep Learning by Dynamic Model Averaging. In Conference Proceedings ECML-PKDD 2018

2 Kamp, M., Boley, M., Keren, D., Schuster, A., Sharfman, I.: Communication-efficient distributed online prediction by dynamic model synchronization. In: Machine Learning and Knowledge Discovery in Databases, Springer (2014), S. 623-639

3 Qadah, E., Mock, M., Alevizos, E., Fuchs, G.: A distributed online learning approach for pattern prediction over movement event streams with apache flink. Big Mobility Data Analytics Workshop on the EDBT/ICT Joint Conference. Wien, 2018

4 McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. 2017, S. 1273-1282

5 Adilova, L., Paul, N., Schlicht, P.: Introducing Noise in Decentralized Training of Neural Networks. ECML PKDD Workshop DMLE 2018

6 Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. Multi- Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. International Conference on Medical Image Computing and Computer Assisted Intervention 2018, Workshop BrainLes

7 Agarwal, N., Suresh, A. T., Xinnan Yu, F., Yu, Kumar, S. & McMahan. B. (2018): cpSGD: Communication-efficient and differentially-private distributed SGD. In: Advances in Neural Information Processing Systems

Bildquellen

Seite 1: © metamorworks - stock.adobe.com