

Innovationsstandort Schweiz, Teil 6/10

Maschinelles Lernen entzaubert: übermenschliche Leistung verstehen

Maschinelles Lernen kann Menschen bei klar definierten Aufgaben übertreffen – aber nur unter den richtigen Bedingungen. Teil 6 der Serie «Innovationsstandort Schweiz» zeigt, warum Datenqualität, Benchmarks und Erklärbarkeit über den erfolgreichen Einsatz der Technologie entscheiden.

› Johan Pocard, Iason Kastanis

Die Nachrichten zum Thema künstliche Intelligenz sind voll von beeindruckenden Behauptungen. Schlagzeilen verkünden Systeme, die scheinbar über übermenschliche Leistungsfähigkeit und menschliches Denkvermögen verfügen – und die für eine Vielzahl von Aufgaben eingesetzt werden können.

Diese Versprechen sind verlockend, werden jedoch häufig missverstanden und führen zu unrealistischen Erwartungen darüber, was heutige Systeme des maschinellen Lernens tatsächlich leisten können. Die Realität ist zwar tatsächlich beeindruckend, aber differenzierter. Modernes maschinelles Lernen kann in bestimmten Situationen in der Tat den Menschen übertreffen, jedoch nur unter wohldefinierten Bedingungen. Ebenso lernen heutige KI-Systeme nicht kontinuierlich aus Erfahrungen, so wie es Menschen tun. Ihre Fähigkeiten hängen von der Aufgabe, den verfügbaren Daten und der Art und Weise ab, wie sie trainiert und bewertet werden. Diese Stärken und ihre Grenzen zu verstehen, ist unerlässlich, um zuverlässige KI-Systeme zu entwickeln und sie erfolgreich in praktischen Anwendungen einzusetzen.

Übermenschliche Leistung

Ist es für eine Maschine tatsächlich möglich übermenschliche Leistungen zu erbringen? Die kurze Antwort lautet: Ja. Dies gilt allerdings nur bei eng gefassten und klar definierten Aufgaben. Wenn Menschen hören, dass ein KI-System übermenschliche Leistungen erbracht hat, stellen sie sich oft eine Maschine vor, die generell intelligenter ist als ein Mensch. In Wirklichkeit zeichnen sich maschinelle

Lernsysteme nur bei den spezifischen Aufgaben aus, für die sie entwickelt und trainiert wurden.

Eine gängige Kategorie des maschinellen Lernens besteht aus Algorithmen, die anhand von annotierten, also zuvor gekennzeichneten Daten trainiert werden. Beispielsweise kann ein Modell im Bereich der Bildverarbeitung lernen, Fehler an einem Fertigungsprodukt zu identifizieren oder Objekte in einem Bild zu erkennen. Da das Modell anhand menschlich bereitgestellter Kennzeichnungen lernt, hat deren Qualität grossen Einfluss auf seine Leistungsfähigkeit. Ist es jedoch einmal trainiert, führt es dieselbe Aufgabe viel schneller, konsistenter und ohne Ermüdung aus. Ein solches System kann Millionen von Bildern in einer Zeit verarbeiten, in der eine menschliche Fachperson nur einen Bruchteil davon prüfen könnte.

Auch bei sequenziellen Entscheidungsaufgaben können Computersysteme übermenschliche Leistungen erzielen. Schach liefert dafür ein bekanntes Beispiel – allerdings auf unterschiedlichen Wegen. Deep Blue setzte vor allem auf enorme Rechenleistung, eine komplexe Bewertungsfunk-



kurz & bündig

- › Datenqualität, Trainingsbedingungen und Einsatzumfeld sind entscheidend, wenn KI menschliche Leistung übertreffen soll.
- › Benchmarks und Erklärbarkeit helfen, die tatsächliche Leistungsfähigkeit von KI-Systemen realistisch zu beurteilen.
- › Die meisten KI-Systeme lernen nicht automatisch weiter, sondern müssen gezielt mit neuen Daten nachtrainiert und erneut validiert werden.



tion und die systematische Suche nach erfolgversprechenden Zügen. Alpha Zero hingegen trainierte durch Millionen von Partien gegen sich selbst ein neuronales Netz, das vielversprechende Züge erkennt und so die Suche gezielt auf aussichtsreiche Stellungen lenkt, statt wie Deep Blue sehr grosse Mengen möglicher Stellungen systematisch durchzurechnen. Beide Ansätze zeigen, dass übermenschliche Leistungen bei klar definierten Aufgaben möglich sind, ohne dass ein System über allgemeine menschliche Intelligenz verfügen muss.

Passende Voraussetzungen

Ob ein System des maschinellen Lernens menschliche Leistungen übertreffen kann, hängt von mehreren Voraussetzungen ab. Das Ziel muss klar definiert und der Erfolg messbar sein, es müssen ausreichend repräsentative Daten oder Simulationen verfügbar sein, und die Umgebung sollte relativ stabil bleiben. Ohne diese Voraussetzungen wird es wesentlich schwieriger, durchwegs übermenschliche Leistungen zu erzielen.

Am Centre Suisse d'Electronique et de Microtechnique (CSEM) werden regelmässig Algorithmen für übermenschliche

Leistungen in einer Vielzahl von Anwendungsbereichen trainiert. So wurde beispielsweise ein Modell entwickelt, das das Holzvolumen eines Stapels nach einem Holzeinschlag allein anhand von Videoaufnahmen des Objekts vorhersagt und dabei eine weitaus höhere Genauigkeit und Geschwindigkeit erreicht als eine menschliche Schätzung. Ein ähnliches Modell zielte darauf ab, in grossem Massstab gebietsfremde Pflanzenarten entlang von Bahnstrecken anhand von Drohnenbildern zu identifizieren. Diese Modelle können einen enormen Mehrwert bieten, doch ihre Entwicklung erfordert erhebliche Vorabinvestitionen in die Generierung der für das Training benötigten Daten.

Benchmarks setzen

Die Behauptung, übermenschliche Leistungen zu erbringen, wirft eine wichtige Frage auf: Woher wissen wir, dass eine Maschine tatsächlich besser abschneidet als Menschen? Bei manchen Problemen, wie zum Beispiel beim Schach, ist die Leistungsmessung unkompliziert, da das Ergebnis eindeutig ist. Viele Anwendungen des maschinellen Lernens umfassen jedoch Aufgaben, bei denen sich selbst Fachpersonen nicht einig sind oder bei

denen die richtige Antwort nicht direkt erfassbar ist. In solchen Situationen sind sorgfältig konzipierte Benchmarks essenziell. Ein Benchmark definiert die Daten, das Bewertungsverfahren und die Leistungskennzahlen, die zum Vergleich verschiedener Modelle herangezogen werden. Er bietet einen gemeinsamen Bezugspunkt, der es Forschenden und Ingenieuren ermöglicht, Fortschritte objektiv zu messen.

Gute Benchmarks sind unerlässlich, weil eine überragende Leistung im Labor nicht zwangsläufig zu Erfolg in der Praxis führt. Ein Modell kann auf einem kuratierten Testdatensatz verblüffende Ergebnisse erzielen, dann aber enttäuschen, wenn es mit sich ändernden Umgebungsbedingungen, unerwarteten Eingaben oder Daten konfrontiert wird, die bereits geringfügig von der Trainingsverteilung abweichen. Grosse Sprachmodelle veranschaulichen diese Herausforderung gut. Sie werden anhand einer breiten Palette von Benchmark-Suiten bewertet, die Mathematik, logisches Denken, Programmierung und Faktenwissen abdecken. Diese Benchmarks ermöglichen nützliche Vergleiche zwischen Modellen, können jedoch die Vielfalt und Unvorhersehbarkeit realer Nutzerinteraktionen nicht vollständig abbilden.

In ähnlicher Weise haben Standard-Datensätze zur Bilderkennung den Fortschritt im Bereich Computer Vision beschleunigt, indem sie faire Vergleiche zwischen Algorithmen ermöglichten. Industrielle Anwendungen erfordern jedoch oft massgeschneiderte Benchmarks, die die in der Praxis anzutreffenden Betriebsbedingungen besser widerspiegeln. Bei der thermischen Analyse konzentrierte sich das CSEM nicht darauf, möglichst viele Beispiele zu sammeln, sondern auf jene Fälle, die für die Anwendung besonders relevant sind. Weniger kritische Szenarien wurden entsprechend geringer gewichtet. Der daraus resultierende Benchmark ermöglichte es, mehrere Modelle unter Bedingungen zu vergleichen, die der beabsichtigten Anwendung sehr nahekamen.

Benchmarks sollten jedoch niemals zum Selbstzweck werden. Modelle können für einen bestimmten Benchmark optimiert werden, während sie beim zugrunde liegenden Problem kaum echte Fortschritte erzielen. Eine gute Bewertung erfordert daher sowohl standardisierte Benchmarks als auch eine Validierung unter realistischen Betriebsbedingungen.

Bedeutung der Erklärbarkeit

Eine hohe Vorhersagegenauigkeit wird oft als oberstes Ziel des maschinellen Lernens angesehen. In der Praxis reicht Genauigkeit allein jedoch selten aus. Nutzer müssen verstehen, warum ein Modell zu einer bestimmten Entscheidung gelangt ist, bevor sie ihm vertrauen können – insbesondere bei Anwendungen, bei denen es um Sicherheit, regulatorische Anforderungen oder erhebliche betriebliche Konsequenzen geht. Erklärbarkeit hilft dabei, die Mechanismen hinter einer Vorhersage aufzudecken, und schafft Vertrauen, dass sich das Modell auf aussagekräftige Informationen stützt und nicht auf zufällige Zusammenhänge. Auch während der Entwicklung spielt die Erklärbarkeit eine wichtige Rolle. Wenn Fachpersonen verstehen, warum ein Modell erfolgreich ist

oder versagt, können sie Schwachstellen in den Trainingsdaten identifizieren, unerwartete Verzerrungen erkennen und das Modell effizienter verbessern. In vielen Fällen spart das Verständnis des Modellverhaltens wesentlich mehr Entwicklungszeit ein, als einfach nur grössere Modelle auszuprobieren oder ziellos zusätzliche Daten zu sammeln.

Überwachung der Modelle

Eine in der Bildverarbeitung weitverbreitete numerische Technik zur Erklärbarkeit ist die Erstellung von Heatmaps. Diese Visualisierungen heben die Bereiche eines Bildes hervor, die am stärksten zur Vorhersage beigetragen haben. Fachpersonen können damit überprüfen, ob sich das Modell auf das Zielobjekt konzentriert oder durch irrelevante Hintergrundmerkmale abgelenkt wird. Über numerische Ansätze hinaus beschleunigt die Kenntnis des zugrunde liegenden physikalischen Prozesses die Entwicklung, indem Probleme direkt identifiziert werden und somit zeitaufwendige Trial-and-Error-Ansätze vermieden werden.

Obwohl Methoden zur Erklärbarkeit die inneren Abläufe komplexer neuronaler Netze nicht vollständig offenlegen können, sind sie zu einem unverzichtbaren Werkzeug für die Fehlerbehebung, die Einhaltung gesetzlicher Vorschriften, die Akzeptanz bei den Nutzenden und die effiziente Modellentwicklung geworden.

Grosse Sprachmodelle, sogenannte Large Language Models (LLMs), haben die Art und Weise, wie Menschen mit KI interagieren, grundlegend verändert. Sie können Fragen beantworten, Dokumente zusammenfassen, Software-Code generieren und bei vielen kreativen und technischen Aufgaben helfen. Trotz dieser beeindruckenden Fähigkeiten sollten sie nicht als autonome Experten betrachtet werden. Im Gegensatz zu herkömmlicher Software generieren LLMs Antworten, indem sie wahrscheinliche Wortfolgen vorhersagen, anstatt gesicherte Fakten

SWISS EXCELLENCE FORUM

Network to Success

Wirkungsorientierte Unternehmensführung

Führung, die wirkt

Führungsseminare mit Fokus auf Wirkung, Transformations- und Zukunftsfähigkeit.



- Operational Excellence: ab November 2026
- Executive Excellence: ab März 2027

KI-Webinar für Führungskräfte

KI gezielt für bessere Entscheidungen, höhere Produktivität und wirksamere Führung einsetzen. Praxisnah und wirkungsvoll.



4 Module, online
ab Februar 2027

SWISS EXCELLENCE FORUM
Kompetenzzentrum für
wirkungs- und zukunftsorientierte
Unternehmensführung



Telefon 041 229 30 40
swiss-excellence-forum.ch

abzurufen. Infolgedessen produzieren sie manchmal «Halluzinationen», sprich, Antworten, die völlig plausibel klingen, aber falsch sind. Noch herausfordernder ist, dass diese Antworten oft mit grosser Überzeugung formuliert werden, was es schwer macht, sie zu erkennen. Die Bewertung der Leistung von LLMs ist zudem von Natur aus schwieriger als die Bewertung vieler herkömmlicher Modelle des maschinellen Lernens. Da ihre Ergebnisse in natürlicher Sprache und nicht als numerische Vorhersagen vorliegen, ist die Richtigkeit oft subjektiv und lässt sich nicht immer mit einer einzigen Metrik quantifizieren. Im Gegensatz zu direkten Quantifizierungsmethoden erfordert die Bewertung der Richtigkeit einer Antwort eines LLMs den jeweiligen Kontext.

Aus diesen Gründen ist es in vielen Anwendungsfällen nach wie vor bewährte Praxis, einen Menschen in den Prozess einzubeziehen. Anstatt einem LLM uneingeschränkte Kontrolle zu gewähren, führen erfolgreiche Systeme häufig interpretierbare Zwischenschichten ein, die die Handlungsmöglichkeiten des Modells einschränken. Ein Beispiel dafür ist ein vom CSEM entwickeltes Laborautomatisierungssystem zur Generierung von Roboterprogrammen unter Verwendung natürlicher Sprache. Anstatt direkt grundlegende Bewegungsbefehle zu generieren, ist ein LLM darauf beschränkt, vorhandene, validierte Aufgabenblöcke zusammenzusetzen, die verifizierte Roboteraktionen enthalten. Diese Zwischenschicht ermöglicht es Fachpersonen, von der Flexibilität der natürlichen Sprache zu profitieren und gleichzeitig Sicherheit, Vorhersehbarkeit und Nachvollziehbarkeit zu gewährleisten.

Stetige Selbstverbesserung?

Ein weitverbreitetes Missverständnis in Bezug auf KI ist, dass sie kontinuierlich lernt. Menschen lernen von Natur aus jeden Tag aus Erfahrungen und passen ihr Wissen schrittweise an, wenn sie mit neuen Situationen konfrontiert werden.

Die meisten eingesetzten «Machine-Learning-Systeme» funktionieren hingegen anders. Nach dem Training bleiben ihre Parameter unverändert, bis technische Fachpersonen das Modell bewusst aktualisieren. Die Verbesserung eines im Einsatz befindlichen Modells ist eher ein strukturierter technischer Prozess als ein automatischer Lernmechanismus. Zunächst müssen neue Daten gesammelt, bereinigt und annotiert werden. Anschliessend wird das Modell neu trainiert, gründlich anhand bestehender Benchmarks getestet, unter realistischen Bedingungen validiert und schliesslich erneut ausgerollt. Dieser Prozess stellt sicher, dass sich die Leistung verbessert, ohne dass unerwartete Fehler auftreten.

Aus eigenen Fehlern lernen

Eine wichtige Quelle für neue Trainingsdaten ist die Aufzeichnung von Situationen, in denen das System schlecht abschneidet. In der Robotik und bei autonomen Systemen werden Fehlerfälle oft protokolliert und in zukünftige Trainingsdatensätze integriert. Techniken wie Datensatz Aggregation (Dagger) ergänzen die Trainingsdaten schrittweise mit Beispielen, die während des Einsatzes gesammelt wurden. Dadurch werden zukünftige Versionen des Modells robuster, ohne die kontrollierten Entwicklungs- und Validierungsprozesse zu umgehen. Obwohl die Forschung zum lebenslangen und kontinuierlichen maschinellen Lernen stetig voranschreitet, setzen die meisten industriellen KI-Systeme aus Stabilitätsgründen nach wie vor auf ein periodisches Umtrainieren statt auf eine kontinuierliche Selbstaktualisierung im Produktionsbetrieb.

Die Leistungsfähigkeit eines Systems des maschinellen Lernens hängt wesentlich von den Daten ab, mit denen es trainiert wurde. Das Sammeln repräsentativer, hochwertiger Trainingsdaten ist oft der zeitaufwendigste Teil eines KI-Projekts. Die Daten sollten die in der Praxis auftretenden Bedingungen genau widerspiegeln, wäh-

rend die zugehörigen Labels so akkurat und konsistent wie möglich sein sollten.

Viele Faktoren können die Modellleistung im Einsatz beeinträchtigen. Die Trainingsdaten können Verzerrungen enthalten, seltene Ereignisse unterrepräsentiert sein oder die Betriebsumgebung kann sich allmählich verändern, was zu einer sogenannten Verteilungsverschiebung führt. Selbst vermeintlich kleine Änderungen können unerwartet grosse Auswirkungen haben. Beispielsweise kann der Austausch eines Kamerasensors die Bildeigenschaften subtil verändern und die Genauigkeit eines Bildverarbeitungsmodells verringern, das zuvor gute Ergebnisse lieferte. Ebenso können sich Änderungen der Lichtverhältnisse erheblich auf die Bildqualität auswirken und zu falschen Vorhersagen führen, wenn solche Bedingungen während des Trainings nicht ausreichend berücksichtigt wurden.

Diese Beispiele verdeutlichen ein grundlegendes Prinzip des maschinellen Lernens: Viele KI-Probleme sind letztlich Datenprobleme. Die Verbesserung der Datenqualität führt oft zu grösseren Gewinnen als die Entwicklung ausgefeilterer Algorithmen. Eine elegante Lösung für das häufige Problem des Datenmangels, die am CSEM zum Einsatz kommt, ist die Generierung synthetischer Datensätze. Anstatt eine grosse Anzahl von Bildern manuell als Beispiele zu erfassen, wird eine Szene mit einer fotorealistischen 3D-Rendering-Software erstellt. Die Software kann ein gewisses Mass an Zufälligkeit einführen und mit geringem Aufwand Tausende unterschiedlicher Beispiele generieren. Mit diesen kann anschliessend ein Modell trainiert werden.

Möglichkeiten und Grenzen

Maschinelle Lernsysteme können zwar übermenschliche Leistungen erzielen, jedoch nur im Rahmen sorgfältig definierter Aufgaben und gestützt auf repräsentative Daten, geeignete Trainingsmethoden sowie umfassende Bewertungs-

verfahren. Zuverlässige Benchmarks sind unerlässlich, um aussagekräftige Fortschritte nachzuweisen, während Erklärbarkeit Entwicklern hilft, das Modellverhalten zu verstehen, das Vertrauen der Nutzer zu stärken und Systeme effizienter zu verbessern. Grosse Sprachmodelle verdeutlichen zudem, dass beeindruckende Fähigkeiten nach wie vor menschliche Überwachung und sorgfältig konzipierte Sicherheitsvorkehrungen erfordern, um zuverlässig zu arbeiten.

Schliesslich verbessern sich die meisten KI-Systeme trotz häufiger Verweise auf kontinuierliches Lernen nach ihrer Einführung nicht automatisch. Ihre Weiterentwicklung hängt von der Erfassung hochwertiger Daten, dem erneuten Training der Modelle und der Validierung der Verbesserungen im Rahmen eines sorgfältig durchgeführten technischen Prozesses ab. Das Verständnis dieser Prinzipien hilft dabei, echte Fortschritte von Marketing-Hype zu unterscheiden, und ermöglicht die Entwicklung von KI-Systemen, die nicht nur leistungsstark, sondern auch zuverlässig, interpretierbar und in realen Anwendungen effektiv sind. «

Dieser Artikel ist Teil einer Serie zum Thema KI, die in Zusammenarbeit mit der Schweizerischen Akademie der Technischen Wissenschaften SATW erarbeitet wird. Die Autorinnen und Autoren sind SAIROP-Partner. Mit der Swiss Artificial Intelligence Research Overview Plattform SAIROP bringt die SATW wichtige Akteure des Schweizer KI-Ökosystems zusammen und bietet einen Überblick über KI-Dienstleistungen, Forschungsprojekte und Kooperationspartner. SAIROP soll die Zusammenarbeit zwischen Forschung, Industrie und Verwaltung fördern und den Wissenstransfer in der Schweiz unterstützen. Die SATW ist das Expertennetzwerk im Bereich Technikwissenschaften in der Schweiz und steht im Kontakt mit den massgeblichen Schweizer Gremien für Wissenschaft, Politik und Industrie. Die SATW ist politisch unabhängig und nicht kommerziell.



Serie: Innovationsstandort Schweiz

04–05/26	Internationale Ansiedlungen als Wachstumstreiber
06/26	Responsible AI als Chance für Schweizer KMU
07–08/26	KI schafft neue Arbeitsplätze und Kompetenzen
07–08/26	Warum Daten den Unterschied machen
09/26	Data Fairness Label
09/26	Maschinelles Lernen entzaubert
10/26	Prompt Engineering: Techniken zum Erstellen von Prompts
11–12/26	Von Prompt Engineering zu Context Engineering
01–02/27	Low-Code-Agenten mit Copilot erstellen
03/27	Expertensysteme und Entscheidungshilfen



Porträt



Johan Pocard

Centre Suisse d'Electronique et de Microtechnique CSEM

Johan Pocard hat einen Master of Science in Maschinenbau von der Eidgenössischen Technischen Hochschule in Lausanne EPFL. Er hat an der Harvard University im Bereich Soft Robotics geforscht und arbeitet seit 2022 beim CSEM im Bereich Industrierobotik und maschinelles Lernen.



Dr. Iason Kastanis

Gruppenleiter, Centre Suisse d'Electronique et de Microtechnique CSEM

Iason Kastanis hat einen Dokortitel in Informatik vom University College London. Seit mehr als einem Jahrzehnt entwickelt er Predictive-Analytics-Lösungen für die Industrie. Er hat Projekte in den Bereichen künstliche

Intelligenz, Computer Vision, Signalverarbeitung und Optimierung geleitet, in denen die entwickelten Technologien produktiv eingesetzt werden. Derzeit leitet er beim CSEM die Gruppe Predictive Analytics.



Kontakt

info@sairop.swiss, johan.pocard@csem.ch, iason.kastanis@csem.ch
www.sairop.swiss
www.csem.ch
www.satw.ch