# TIEFE NETZE

TIEFE NETZE

# VON MASCHINEN LERNEN

ULLRICH KÖTHE

**Während der Operation eines Patienten berechnen sogenannte invertierbare neuronale Netze den Sauerstoffgehalt des Blutes in Geweben und Gefäßen. Dabei erkennen diese Netze automatisch Bildbereiche, die von Instrumenten oder Tüchern verdeckt sind, als „kein Gewebe“ und ignorieren sie. Ein Beispiel für eine medizinische Anwendung, die Forschungsergebnisse der Arbeitsgruppe „Explainable Machine Learning“ am Interdisziplinären Zentrum für Wissenschaftliches Rechnen der Universität Heidelberg nutzt. Das maschinelle Lernen bildet gemeinsam mit der logikbasierten KI und der Mustererkennung die drei grundlegenden Ansätze auf dem Gebiet der Künstlichen Intelligenz (KI). Die Anwendungsgebiete der KI reichen vom Sieg eines Computers über die weltbesten menschlichen Go-Spieler bis zu Bildsegmentierung, Fluoreszenzmikroskopie oder Diagnostik in der Krebsmedizin oder der Augenheilkunde.**

Nahezu jedem ist heute im Alltag einer der größten Erfolge der Künstlichen Intelligenz (KI) vertraut: das Navigationsprogramm. Es speichert relevante geographische Informationen in einer riesigen Datenbank und beantwortet mittels effizienter Suchverfahren Anfragen nach dem kürzesten Weg oder der schnellsten Verbindung. Das Navigationsprogramm ist ein Anwendungsbeispiel für die „logikbasierte KI", einen von drei grundlegenden Ansätzen, die sich im Feld der Künstlichen Intelligenz unterscheiden lassen. Neben der logikbasierte KI gibt es die „Mustererkennung" und das „maschinelle Lernen". Alle drei Ansätze existierten mehr oder weniger von Anfang an nebeneinander. Das „Perzeptron" beispielsweise – eines der ersten künstlichen neuronalen Netze – wurde bereits im Jahr 1958 von dem US-amerikanischen Wissenschaftler Frank Rosenblatt eingeführt. Geändert hat sich hingegen, welcher Ansatz jeweils im Mittelpunkt des Interesses stand: Bis etwa zum Jahr 1985 war es die logikbasierte KI, zwischen 1985 und 2012 die Mustererkennung – seit 2012 hat das maschinelle Lernen seinen Siegeszug angetreten. Mit dem maschinellen Lernen und wie man es optimieren kann, beschäftigt sich auch unsere Forschergruppe im Rahmen des Exzellenzclusters „STRUKTUREN" der Universität Heidelberg.

**Die Logik als Basis**

Die logikbasierte KI entstand aus dem konsequenten Weiterdenken der Erfindung des Computers um 1940. Es hatte sich seinerzeit schnell gezeigt, dass Computer nicht nur einfache Berechnungen automatisieren können wie etwa ein Taschenrechner. Sie können auch anspruchsvollere Aufgaben bewältigen, beispielsweise Differenzialgleichungen lösen oder große Datenmengen systematisch nach relevanten Informationen durchsuchen. Digitale Computer beruhen auf der mathematischen Formalisierung der Logik, von George Boole (1815 bis 1864) bis zu Alan Turing (1912 bis 1954). Der Erfolg dieses Vorgehens legte die Hypothese nahe, dass geistige Leistungen generell der Ausdruck eines ausgeklügelten Anwendens logischer Verfahren sind.

Die logikbasierte KI erforscht, wie sämtliches Wissen über eine Domäne der Welt formal in Wissensbasen gespeichert werden kann und wie daraus mittels logischer Schlussregeln neues Wissen und konkrete Handlungsempfehlungen abgeleitet werden können. Nicht nur zur eingangs erwähnten Navigation, auch in den Naturwissenschaften ist die logikbasierte KI unverzichtbar: Computer-Algebraprogramme wie „Mathematica" und „Maple" erledigen in wenigen Sekunden schwierige mathematische Herleitungen, für die ein Experte Stunden oder gar Tage benötigt – und das, ohne sich dabei auf halbem Wege zu verrechnen.

**Das Interdisziplinäre Zentrum für Wissenschaftliches Rechnen**

Das Interdisziplinäre Zentrum für Wissenschaftliches Rechnen (IWR) wurde 1987 als bundesweit erstes universitäres Forschungszentrum seiner Art gegründet. Die Forschung am IWR befasst sich mit Fragestellungen aus Natur-, Technik- und Geisteswissenschaften und bearbeitet sie mit dem Methodenrepertoire des Wissenschaftlichen Rechnens: der mathematischen Modellierung, Simulation und Optimierung, der Bild- und Datenverarbeitung sowie der Visualisierung. Als Querschnittsdisziplin trägt das Wissenschaftliche Rechnen entscheidend zur Lösung anspruchsvoller Probleme aus Wissenschaft und Technik bei und gilt damit als eine Schlüsseltechnologie des 21. Jahrhunderts. Seine Methoden kommen bei so unterschiedlichen Fragestellungen zum Einsatz wie dem Entwurf effizienter Brennstoffzellen, der Simulation der Vorgänge beim Hirninfarkt, der Prognose des Pestizidabbaus im Boden oder auch der Optimierung von Bewegungsabläufen.

Das IWR umfasst heute 50 Forscherteams aus unterschiedlichen Fakultäten sowie neun von jungen Wissenschaftlern geführte Nachwuchsgruppen. Rund 500 Forscherinnen und Forscher arbeiten im Rahmen des Zentrums in interdisziplinären Kooperationen zusammen. Neben Mathematik, Physik, Chemie und Informatik sowie den Lebenswissenschaften sind zunehmend auch Wirtschafts- und Sozialwissenschaften, Psychologie, Kognitionswissenschaften sowie Geistes- und Kulturwissenschaften vertreten. Die Infrastruktur des IWR umfasst unter anderem Hochleistungsrechner, 3D-Graphiklabore sowie spezielle Laser-Scanner. Auf Initiative des IWR entstand 2007 die im Rahmen der Exzellenzinitiative geförderte „Heidelberger Graduiertenschule der mathematischen und computergestützten Methoden für die Wissenschaften" (HGS MathComp). Dort forschen derzeit rund 100 Doktorandinnen und Doktoranden aus allen am IWR vertretenen Fächern.

www.iwr.uni-heidelberg.de

Um das Jahr 1980 wurden zwei Nachteile des logikbasierte Ansatzes deutlich. Erstens ist er für viele Fragestellungen nicht effizient genug, weil zahlreiche Verfahren „NP-hart" sind: Das bedeutet, dass ihre Laufzeit exponentiell mit der Problemgröße wächst. Noch schwerer wiegt der zweite Nachteil: Die Logik hat es sehr schwer, mit Unsicherheit umzugehen. Logische Aussagen sind entweder wahr oder falsch – die reale Welt aber lässt sich nicht in ein solch einfaches Schwarz-Weiß-Schema zwängen. Im Leben gilt es, ständig Entscheidungen zu treffen, ohne zuvor das gesamte verfügbare Wissen sammeln und auswerten zu können; auch der beste Plan wird häufig aufgrund zufälliger Ereignisse hinfällig. Das gilt selbst in exakten Wissenschaften wie der Physik. Dort sind der möglichen Genauigkeit von Messungen und Vorhersagen prinzipielle Grenzen gesetzt, beispielsweise in chaotischen Systemen wie dem Wetter oder bei der Unschärferelation in der Quantenmechanik. Der zweite Ansatz der Künstlichen Intelligenz – die Mustererkennung – postuliert deshalb als zentrale Entitäten der Künstlichen Intelligenz Muster anstelle von Fakten und Regeln.

**Das Erkennen von Mustern**

Die Mustererkennung wurde besonders von der Wahrnehmungs- und Gestaltpsychologie inspiriert. Deren Kernaussage, „Das Ganze ist mehr als die Summe seiner Teile", verdeutlicht, dass komplexe Phänomene nur im Zusammenspiel ihrer Teile verstanden werden können. Der Begriff „Muster" wird dabei sehr weit gefasst. Das Elektrokardiogramm eines gesunden und eines herzkranken Menschen etwa weist unterschiedliche zeitliche Muster auf, die dem Arzt bei der Diagnose helfen; in Bildern werden bestimmte Pixelmuster vom Auge des Betrachters als Gesichter, Bäume, Menschen, Tiere oder Fahrzeuge interpretiert.

Als Formalismus, der mathematisch adäquat ist, um derartige Muster zu beschreiben, hat sich die Wahrscheinlichkeitsrechnung erwiesen, insbesondere die „Bayes'sche Statistik". Sie erlaubt präzise Aussagen darüber, wie sich Unsicherheiten und Fehler in den Prämissen und Beobachtungen auf die Unsicherheit und Verlässlichkeit der Folgerungen und Vorhersagen auswirken. Die Frage, ob ein Bild einen Menschen zeigt, muss nun nicht mehr mit einem einfachen Ja oder Nein beantwortet werden – sie kann Unsicherheit ausdrücken, etwa, wenn es sich bei der Gestalt auf dem Bild auch um eine Puppe handeln könnte.

Im Alltag sind die Errungenschaften der Mustererkennung ebenfalls weitverbreitet. Jedem vertraut ist die Gesichtserkennung in digitalen Kameras: Sie sorgt dafür, dass Gesichter möglichst genau fokussiert werden. Die Anwendung mit der momentan größten Relevanz jedoch sind die Internet-Suchmaschinen. Einst gab ein Mustererkennungsalgorithmus

**„Die Anwendung der Künstlichen Intelligenz mit der momentan größten Bedeutung sind die Internet-Suchmaschinen."**

# „Tiefe neuronale Netze erkennen komplexe Muster auch bei sehr großer Variabilität der Erscheinungs-formen zuverlässig und robust.“

namens „PageRank“ den Anstoß zur Gründung der Firma Google – das aktuelle Suchverfahren von Google ist leider geheim. Die Bedeutung einer umfassenden Volltextsuche für die Wissenschaft kann gar nicht hoch genug bewertet werden: Aus Stunden und Tagen in der Bibliothek, um relevante Literatur inklusive ihrer Querverbindungen und Zitate zu recherchieren, sind heute Minuten oder gar Sekunden am Bildschirm geworden.

Doch selbst die Mustererkennung kann viele KI-Probleme nicht lösen: Die verwendeten statistischen Methoden sind zu schwach für die gewaltige Vielfalt real vorkommender Muster. Das wird besonders deutlich, wenn man den Stand der automatischen Übersetzungsprogramme vor rund zehn Jahren betrachtet. Damals wurde der englische Satz „Air travel is known to be the fastest mode of transportation to reach a particular destination“ von einem typischen Algorithmus übersetzt in: „Luftreisen ist bekannt, um die schnellste Weise des Transports zu sein, um einen besonderen Bestimmungsort zu erreichen.“ Noch schlimmer erging es dem Satz „Out of Dell ink for your Dell printer?“. Aus ihm wurde „Aus dem Kleinen engen Tal schwärzen für Ihren Drucker des Kleinen engen Tales mit Tinte?“ – der Algorithmus war nicht imstande, das Wort „Dell“ als Firmenname zu erkennen.

Die breite Einführung des maschinellen Lernens – insbesondere des sogenannten tiefen Lernens – ab dem Jahr 2012 hat die Situation grundlegend verändert. Aktuelle Programme übersetzen auf der Basis neuronaler Netze die Beispielsätze korrekt mit „Der Flugverkehr ist bekanntlich das schnellste Transportmittel, um ein bestimmtes Ziel zu erreichen“ und „Keine Dell-Tinte für Ihren Dell-Drucker?“. Nach wie vor beruht das maschinelle Lernen auf dem Erkennen von Mustern. Es gibt aber mittlerweile einen entscheidenden Unterschied: Zuvor wurden die Modelle, die vorkommende Muster und deren Wahrscheinlichkeiten beschreiben, manuell von Experten erstellt. Heute werden sie mittels geeigneter Trainingsdaten automatisch optimiert, also maschinell „gelernt“.

Überraschenderweise hat sich gezeigt, dass maschinell gelernte Modelle den manuell erstellten deutlich überlegen sind. Dafür gibt es mehrere Gründe: Erstens haben sich die Lernverfahren in den letzten Jahren mit der Entwicklung schneller Hardware (insbesondere der GPUs, der „Graphics Processing Units“) und neuer Algorithmen stark verbessert. Zweitens sind inzwischen sehr große Trainingsdatensätze verfügbar, die ein Mensch nicht mehr umfassend verwerten kann – ein Computer aber sehr wohl. Drittens erweisen sich moderne neuronale Netze mit bis zu 1.000 Neuronenschichten – sogenannte „tiefe Netze“ – als extrem mächtige Bayes’sche Modelle. Sie erkennen komplexe Muster auch bei sehr großer Variabilität der Erscheinungsformen zuverlässig und robust, beispielsweise Personen, Fahrzeuge und Hindernisse in natürlichen Bildern. Die Vorreiter der künstlichen neuronalen Netze – Geoffrey Hinton, Yoshua Bengio und Yann LeCun – wurden für ihre Arbeiten im Jahre 2018 mit dem Turing Award ausgezeichnet, dem „Nobelpreis“ der Informatik. Die von ihnen und vielen anderen Wissenschaftlern initiierten Methoden eröffnen auch für die Forschung in den Natur- und Lebenswissenschaften völlig neue Möglichkeiten.

**Biologische Anwendungen**

Tiefe Netze werden beispielsweise genutzt, um bildgebende Verfahren zu verbessern. Exemplarisch zeigen das zwei Anwendungen der Fluoreszenzmikroskopie, einer modernen mikroskopischen Methode. Dabei werden bestimmte Strukturen im Innern von Zellen gezielt mit sogenannten Fluorophoren gefärbt, so dass genau diese zellulären Strukturen aufleuchten, wenn sie mit geeignetem Laserlicht dazu angeregt werden. Für eine optimale Bildqualität sollte der Anregungslaser eine hohe Intensität haben. Dies schädigt aber die Zellen (Phototoxizität) und verhindert, sie in ihrer Entwicklung zu beobachten. Niedrige Laserintensitäten wiederum führen zu stark verrauschten Bildern. Die Lösung, die Forscher der Max-Planck-Gesellschaft um Florian Jug und Eugene Myers dafür gefunden haben: Man nimmt in einer Trainingsphase Bildpaare auf, die die gleichen Zellen zunächst mit niedriger und dann mit hoher Intensität abbilden. Anhand dieser Daten erlernt ein neuronales Netz, Bilder hoher Qualität aus den korrespondierenden verrauschten Bildern zu rekonstruieren. Eine solche „content-aware image restoration“ macht es nach einer Trainingsphase möglich, niedrige Laserintensitäten zu benutzen, ohne dabei starke Abstriche bei der Bildqualität hinnehmen zu müssen.

Eine andere Variante der Fluoreszenzmikroskopie, für die Stefan Hell, Eric Betzig und William Moerner im Jahr 2014 den Chemie-Nobelpreis erhielten, versucht, das Abbe’sche Auflösungslimit von 200 Nanometern zu überwinden. Hierzu sorgt man beispielsweise mit chemischen Tricks dafür, dass zu jedem Zeitpunkt nur wenige Fluorophore aktiv sind. Deren Lichtpunkte sind dann im Bild so weit separiert, dass man ihre Position auf 1/10 Pixel, also 20 Nanometer, genau lokalisieren kann. Um ein superauflösendes Bild zusammenzusetzen, braucht man allerdings bis zu 10.000 Einzelaufnahmen. Mit nur 100 Einzelaufnahmen kommt man aus, wenn viele Fluorophore gleichzeitig aktiviert werden. Weil jetzt aber die Lichtpunkte überlappen, versagen klassische Verfahren zur Lokalisierung. Kürzlich ist es dem Team von Anna Kreshuk am European Molecular Biology Laboratory (EMBL) mit neuronalen Netzen gelungen, auch bei hoher Punktdichte eine zuverlässige und genaue Lokalisation zu gewährleisten. Praktisch bedeutet das: Die Aufnahme eines superauflösenden Bildes dauert nicht

# „Neuronale Netze erreichen in der Medizin bei vielen diagnostischen Fragen die Genauigkeit der besten Experten.“

mehrere Minuten, sondern nur noch wenige Sekunden – damit können auch zeitliche Veränderungen der Zellen sichtbar gemacht werden.

Eine noch größere Bedeutung hat das tiefe Lernen für die Bildsegmentierung und -klassifikation. Ein spektakuläres Beispiel ist der aus einer Kooperation von Janelia Research und Google unter Leitung von Stephen Plaza und Viren Jain hervorgegangene „Hemibrain-Datensatz“ für die Hirnforschung. Er umfasst einen großen Teil des Gehirns einer Fruchtfliege (Drosophila melanogaster) mit dem Ziel, das „Konnektom“ zu verstehen, das vollständige Netzwerk aller Nervenzellen (Neuronen). Um gleichzeitig die feinsten neuronalen Strukturen und die volle Ausdehnung der Nervenzellfortsätze abbilden zu können, benötigt man eine dreidimensionale Auflösung von acht Nanometern und ein Bildfeld von 0,25 Millimetern in jeder Raumrichtung, also circa $30.000^3$ Pixel. Die Segmentierung sämtlicher 25.000 Nervenzellen und 20 Millionen Synapsen – der Verknüpfungen der Nervenzellen – in diesem 27 Terabyte umfassenden Datensatz war nur durch neue Verfahren des tiefen Lernens in Verbindung mit einer rigorosen Qualitätskontrolle durch menschliche Experten („proof reading“) möglich.

Ähnliche Entwicklungen beobachten wir in anderen Bereichen der Biologie, beispielsweise bei der Hochdurchsatzmikroskopie. In der Medizin wurde unlängst gezeigt, dass neuronale Netze bei vielen diagnostischen Fragen, vor allem in der Krebsmedizin und der Augenheilkunde, die Genauigkeit der besten Experten erreichen. Daher etabliert sich zunehmend ein effizientes arbeitsteiliges Verfahren: Der Computer weist den Arzt auf interessante Bereiche in den Daten hin und schlägt Interpretationen vor – der Arzt konzentriert sich auf Diagnose und Bewertung.

**Induktives und transduktives Vorgehen**

Gerade die Erfolge in der Medizin verdeutlichen allerdings auch ein prinzipielles Problem aktueller neuronaler Netze: Sie liefern zwar hervorragende Ergebnisse, aber es ist nicht offensichtlich, wie sie dabei vorgehen und auf welche Muster sie ihre Entscheidungen stützen. Ein solches Black-Box-Verhalten ist für Anwendungen inakzeptabel, bei denen Fehler potenziell lebensbedrohlich sind. Hier hilft wiederum ein Blick in die Geschichte, um die Ursachen des Problems zu verstehen.

Die Naturwissenschaften verfolgen traditionell einen induktiven Ansatz: Daten aus Beobachtungen und Experimenten werden benutzt, um dahinterliegende allgemeine Gesetze zu identifizieren. Dieses theoretische Verständnis ermöglicht es, korrekte Vorhersagen für neue Experimente und Situationen zu treffen. Das maschinelle Lernen ist anfangs ebenso vorgegangen. Die Erfolge aber waren bis circa 1990 bescheiden: Die meisten Fragestellungen waren zu komplex für die damaligen theoretischen Werkzeuge. Der russisch-amerikanische Mathematiker Vladimir Vapnik hat deshalb den „transduktiven Ansatz“ eingeführt, der die Aufgabe drastisch vereinfacht, indem er Vorhersagen direkt von Beobachtungen ableitet. Das „inverse Pendel“ verdeutlicht den Unterschied zwischen einem induktiven und einem transduktiven Vorgehen. Bei diesem Beispiel muss ein aufrechter Stab auf einem Finger oder Roboterwagen balanciert werden. Der induktive Ansatz stellt zunächst die physikalischen Bewegungsgleichungen für jedes beliebige inverse Pendel auf. Daraus lassen sich – nach Einsetzen

DEEP NETWORKS

# LEARNING FROM MACHINES

ULLRICH KÖTHE

Artificial intelligence (AI), a field of study that is barely 70 years old, has already made major inroads into daily life, for example in the form of navigation programmes, internet search engines and automatic language translation. The widespread adoption of neural networks and deep learning over the last decade has led to substantial progress in the quality of results and the scope of potential applications. The present article illustrates this with a description of recent advances in biological and medical imaging.

Heidelberg University, and specifically my research group “Explainable Machine Learning”, focuses on artificial intelligence applications in the natural and life sciences. These applications pose additional challenges to AI methods: models should not merely compute accurate results and good predictions, they also need to make their reasoning explicit and provide insights into the underlying natural phenomena. Current AI algorithms, which mainly follow the so-called transductive approach, do not offer this capability. We aim at designing new model types that open up the black box by combining the transductive approach with the classical inductive paradigm of scientific inquiry. Promising first examples demonstrate the potential of this combination for interpretable artificial intelligence. ●

**ADJUNCT PROF. ULLRICH KÖTHE** has been heading the research group "Explainable Machine Learning" in Heidelberg University's Visual Learning Lab since 2017. Prior to that, he designed machine learning methods for image analysis in the life sciences in the "Image Analysis and Learning" research group at Heidelberg University's Interdisciplinary Center for Scientific Computing (IWR) and investigated the fundamental performance limits of generic image analysis methods at the University of Hamburg. He was also a leading contributor to several open source software projects for image analysis. Ullrich Köthe's research focuses on artificial neural networks whose processes can be understood and interpreted by humans, and examines how these networks can be used as research tools in the natural and life sciences. In his function as e-learning representative of the computer science department, he also coordinates online teaching in this field.

**Contact: ullrich.koethe@iwr.uni-heidelberg.de**

# "Artificial neural networks in medical applications can answer many diagnostic questions with the same accuracy as the best experts."

**APL. PROF. DR. ULLRICH KÖTHE** leitet seit 2017 die Arbeitsgruppe „Explainable Machine Learning“ im Visual Learning Lab der Universität Heidelberg. Zuvor entwickelte er in der „Image Analysis and Learning“-Gruppe am Interdisziplinären Zentrum für Wissenschaftliches Rechnen (IWR) der Universität Heidelberg maschinelle Lernverfahren für die Bildanalyse in den Lebenswissenschaften und untersuchte an der Universität Hamburg die prinzipiellen Leistungsgrenzen generischer Bildanalyseverfahren. Außerdem war er federführend an mehreren Open-Source-Softwareprojekten für Bildanalysealgorithmen beteiligt. Ullrich Köthe erforscht neuronale Netze, deren Vorgehen für Menschen nachvollziehbar und interpretierbar ist, und wie sich solche Netze als Erkenntniswerkzeuge in den Natur- und Lebenswissenschaften einsetzen lassen. Er koordiniert außerdem als eLearning-Beauftragter der Informatik die Online-Lehre in diesem Bereich.

**Kontakt: ullrich.koethe@iwr.uni-heidelberg.de**

der Parameter eines konkreten Pendels – die passenden Steuerkommandos für die Hand oder den Roboter berechnen. Ein Mensch hingegen lernt das Balancieren eher transduktiv: Durch fortgesetztes Üben setzt das Gehirn die beobachteten Bewegungen des Stabs direkt in die notwendigen Bewegungen der Hand um.

Die Unterscheidung zwischen induktivem und transduktivem Vorgehen ist im maschinellen Lernen weitverbreitet. Die meisten Erfolge der letzten Jahre – von der Bildsegmentierung bis zum Gewinn des Computers gegen die weltbesten menschlichen Go-Spieler – beruhen auf dem transduktiven Ansatz. Dadurch gingen allerdings das theoretische Verständnis und die Interpretierbarkeit der Lösungen verloren. Es ist deshalb an der Zeit, dem induktiven Ansatz wieder mehr Aufmerksamkeit zu schenken. Diesem Ziel widmet sich unsere Arbeitsgruppe „Explainable Machine Learning“ am Interdisziplinären Zentrum für Wissenschaftliches Rechnen (IWR) der Universität Heidelberg. Dazu kooperieren wir im Rahmen des Exzellenzclusters „STRUKTUREN: Emergenz in Natur, Mathematik und komplexen Daten“, des von der Klaus Tschira Stiftung geförderten Informatics4Life-Projekts und anderer Initiativen mit vielen Forschern aus den Natur- und Lebenswissenschaften.

### Invertierbare neuronale Netze

Im Zentrum unserer Forschung stehen sogenannte invertierbare neuronale Netze (INNs). Klassische „feed-forward Netze“ funktionieren nur in einer Richtung, von den Eingaben zu den Ausgaben. Invertierbare neuronale Netze hingegen können in beiden Richtungen betrieben werden – von Eingaben zu Ausgaben und umgekehrt. Wir erforschen, wie man diese Fähigkeit von INNs nutzen kann, um die Vorteile von transduktiven und induktiven Methoden im selben Modell zu vereinen: Der transduktive Aspekt stellt dabei sicher, dass unsere Netze ebenso präzise Ergebnisse liefern wie Standardnetze; der induktive Aspekt sorgt für die Interpretierbarkeit dieser Ergebnisse.

Dabei beachten wir beispielsweise auch die Frage, ob eine gegebene Eingabe überhaupt zur betrachteten Domäne passt, was die Arbeitsgruppe der Informatikerin Lena Maier-Hein am Deutschen Krebsforschungszentrum (DKFZ) in Heidelberg für eine medizinische Anwendung nutzt: Invertierbare neuronale Netze berechnen während der Operation eines Patienten den Sauerstoffgehalt des Blutes in Geweben und Gefäßen aus multispektralen Bildern. Bildbereiche, die während der Operation etwa durch Instrumente oder Tücher verdeckt sind, werden von INNs automatisch als „kein Gewebe“ erkannt und ignoriert. Standardnetze hingegen geben in solchen Bereichen oft unsinnige Antworten. In einer weiteren Arbeit konnten wir zeigen, dass invertierbare Netze unter bestimmten Bedingungen neue Variablen, sogenannte Konzepte, entdecken, die für Menschen intuitiv verständlich sind. Wir finden beispielsweise im sogenannten MNIST-Datensatz rund 20 anschauliche Konzepte, mit denen sich die Variationsbreite handgeschriebener Zahlen nachvollziehbar erklären lässt.

Derartige Konzepte sollen im nächsten Schritt die Basis für kausale Erklärungen liefern. Der Vorteil von kausalen Interpretationen gegenüber rein statistischen Korrelationen liegt auf der Hand. Wenn man keine kontrolliert randomisierten Experimente durchführen kann, ist die datengetriebene Extraktion kausaler Relationen jedoch sehr anspruchsvoll – das haben die Untersuchungen des US-amerikanischen Informatikers Judea Pearl und vieler anderer Wissenschaftler gezeigt. Ein erster Hinweis, was Lernverfahren hier leisten können, wurde vor wenigen Monaten von Forschern des Massachusetts Institute for Technology (MIT) um James Collins publiziert: Ihr Modell war nach Training auf einigen Tausend Molekülen mit bekannter Wirkung in der Lage, aus 100 Millionen Kandidatenmolekülen 23 potenzielle neue Antibiotika zu identifizieren – mehrere von ihnen haben sich in Untersuchungen mit Tieren tatsächlich als hochwirksam erwiesen. Mächtige interpretierbare Modelle, bei denen tiefe Netzwerke automatisch aussagekräftige Variablen und kausale Zusammenhänge identifizieren, sind deshalb ein vielversprechender Weg, um maschinelles Lernen zukünftig – über die Datenanalyse hinaus – als eine allgemeine Erkenntnismethode in den Natur- und Lebenswissenschaften zu etablieren. ●