

Zeitschrift: Horizonte : Schweizer Forschungsmagazin
Herausgeber: Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung
Band: 31 [i.e. 30] (2018)
Heft: 116

Artikel: Die blinden Flecken neuronaler Netze
Autor: Titz, Sven
DOI: <https://doi.org/10.5169/seals-821344>

Nutzungsbedingungen

Die ETH-Bibliothek ist die Anbieterin der digitalisierten Zeitschriften auf E-Periodica. Sie besitzt keine Urheberrechte an den Zeitschriften und ist nicht verantwortlich für deren Inhalte. Die Rechte liegen in der Regel bei den Herausgebern beziehungsweise den externen Rechteinhabern. Das Veröffentlichen von Bildern in Print- und Online-Publikationen sowie auf Social Media-Kanälen oder Webseiten ist nur mit vorheriger Genehmigung der Rechteinhaber erlaubt. [Mehr erfahren](#)

Conditions d'utilisation

L'ETH Library est le fournisseur des revues numérisées. Elle ne détient aucun droit d'auteur sur les revues et n'est pas responsable de leur contenu. En règle générale, les droits sont détenus par les éditeurs ou les détenteurs de droits externes. La reproduction d'images dans des publications imprimées ou en ligne ainsi que sur des canaux de médias sociaux ou des sites web n'est autorisée qu'avec l'accord préalable des détenteurs des droits. [En savoir plus](#)

Terms of use

The ETH Library is the provider of the digitised journals. It does not own any copyrights to the journals and is not responsible for their content. The rights usually lie with the publishers or the external rights holders. Publishing images in print and online publications, as well as on social media channels or websites, is only permitted with the prior consent of the rights holders. [Find out more](#)

Download PDF: 19.02.2026

ETH-Bibliothek Zürich, E-Periodica, <https://www.e-periodica.ch>



Indian elephant

Die blinden Flecken neuronaler Netze

Künstliche Intelligenz vollbringt wahre Kunststücke. Wie sie genau funktioniert, das durchschaut bisher niemand. Nun führen Forschende Algorithmen absichtlich in die Irre, um deren Grenzen zu testen und sie besser zu verstehen. *Von Sven Titz*

Plötzlich hegen einige Fachleute für künstliche Intelligenz (AI) ungewöhnliche Zweifel: «Das maschinelle Lernen ist zur Alchemie geworden», unkte Ali Rahimi von Google neulich in einem Vortrag. Seine Provokation löste eine lebhafte Debatte aus. Rahimi hatte einen Nerv getroffen.

Vielleicht war ein Rückschlag überfällig. In den letzten Jahren haben tiefe neuronale Netze - lernfähige Rechengebilde, die aus mehreren Schichten virtueller Neuronen bestehen - erstaunliche Erfolge gefeiert, etwa in der Sprach- und Bilderkennung. Jetzt folgt das Unbehagen: Weiss man wirklich, was im Innern neuronaler Netze vor

sich geht? Lassen sich die neuen Techniken austricksen? Sind sie ein Sicherheitsrisiko? Diesen Fragen widmen sich neue Forschungsgebiete, die sich «Explainable AI» oder «AI neuroscience» nennen.

Tiefe neuronale Netze (DNN für «deep neural networks») lassen sich auf vielfältige Weise täuschen, wie mehrere Forschende gezeigt haben. Anh Nguyen von der Auburn University zum Beispiel konstruierte Bilder, die für Menschen nicht den geringsten Sinn ergeben, die DNN zur Bilderkennung aber eindeutig als Darstellungen bestimmter Tiere identifizierten.

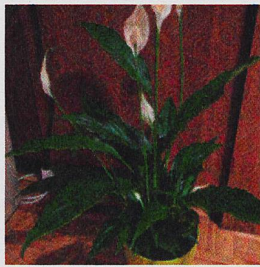
Noch tückischer sind die sogenannten «adversarial» (feindlichen) Testbeispiele.

Realistisch aussehende Bilder werden dabei minimal verändert. Das menschliche Auge nimmt den Unterschied praktisch nicht wahr. Dennoch identifiziert das DNN im manipulierten Bild einen völlig anderen Gegenstand. Der Gruppe von Pascal Frossard von der EPFL gelang es zum Beispiel, dass eine abgebildete Socke für einen Elefanten gehalten wurde.

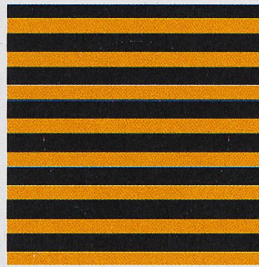
«Systeme auf der Basis von DNN sind derzeit ziemlich verletzlich gegenüber Veränderungen der zugrundeliegenden Daten», sagt Frossard. «Oft können wir keine Garantie für ihre Leistung aussprechen.» Bei Anwendungen im Bereich von Medizin und Sicherheit kann das zu einem echten

Verwirrte neuronale Netzwerke:
Eine Socke wird zum Elefanten,
ein paar Linien zum Schulbus. Mit
diesen Bildern haben Forschende
neuronale Netzwerke getäuscht.

Bilder: S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi
and P. Frossard Proceedings of IEEE CVPR,
2017 (indian elephant, macaw); Nguyen et
al., «DNNs are Easily Fooled», CVPR 2015
(school bus, comic book)



macaw



school bus



comic book

Problem werden. Selbstfahrende Autos
zum Beispiel müssen Verkehrszeichen ver-
lässlich erkennen. Sie dürfen sich durch
Manipulationen nicht täuschen lassen.

Achtbeinige Zebras

Allmählich beginnen Forschende zu ver-
stehen, wie es zu den Fehlern kommt. Ein
Grund ist, dass die Programme mit einer be-
grenzten Menge an Beispieldaten trainiert
werden. Werden sie dann mit ganz anderen
Fällen konfrontiert, geht das gelegentlich
schief. Ein weiterer Grund für das Versagen
ist die Tatsache, dass DNN nicht die struktu-
rell korrekte Wiedergabe von Objekten ler-
nen. «Ein echtes Bild eines vierbeinigen Ze-
bras wird als Zebra klassifiziert», erläutert
Nguyen. «Fügt man dem Zebra im Bild aber
weitere Beine hinzu, ist das DNN eher noch
sicherer, dass es sich um ein Zebra handelt -
selbst wenn das Tier acht Beine hat.»

«Oft können wir keine
Garantie für die Leistung von
tiefen neuronalen Netzen
aussprechen.»

Pascal Frossard

Das Problem: Die DNN ignorieren den
Gesamtaufbau der Bilder. Vielmehr basiert
die Erkennung auf Farb- und Formdetails.
Das ergibt sich jedenfalls aus den ersten
Studien, in denen ermittelt wurde, wie die
DNN im Innern ticken.

Um den Geheimnissen der neuronalen
Netze auf die Schliche zu kommen, nutzen
Nguyen und andere Forschende unter an-
derem Techniken zur Visualisierung. Sie
markieren, welche virtuellen Neuronen auf
welche Eigenschaften von Bildern reagie-
ren. Eines der Resultate: Generell lernen
die ersten Schichten von DNN die Grund-
eigenschaften der Trainingsdaten, wie
Nguyen erläutert. Das sind bei Bildern zum
Beispiel Farben und Linien. Je tiefer man in
ein neuronales Netz vordringt, desto mehr
werden die bereits erfassten Informatio-
nen kombiniert. Die zweite Schicht erfasst
schon Konturen und Schatten. Im Verbund
des Netzes gelingt schliesslich die Erken-
nung von Objekten.

Dabei gibt es erstaunliche Parallelen
zu den Neurowissenschaften: So konnten
Hinweise darauf gefunden werden, dass
einzelne Neuronen im Hirn auf bestimm-
te prominente Personen spezialisiert sein
könnten. Ähnliche Resultate ergaben sich
auch bei den DNN.

Man versucht das Innenleben neuro-
naler Netze auch auf theoretischem Weg
zu entschlüsseln. «Dabei geht es zum Bei-
spiel um mathematische Eigenschaften
der Algorithmen», erklärt Frossard. «Ent-
scheidungsgrenzen» repräsentieren die
Grenzen zwischen verschiedenen Bild-
kategorien. Zum Beispiel wird markiert, ob
ein Bild in die Kategorie «Äpfel» oder die
Kategorie «Birnen» fällt.

Was die Funktionsweise angeht, seien
generell noch viele Fragen offen, sagt Yan-
nic Kilcher vom Data Analytics Lab der ETH
Zürich. Das betrifft die Fehler ebenso wie
das Wunder des Gelingens. Oft liefert selbst
ein Programm, das auf unbekannte Daten
angewandt wird, vernünftige Ergebnisse.
«Warum die neuronalen Netze zu dieser
Verallgemeinerung fähig sind, verstehen
wir noch nicht vollständig», so Kilcher.

Schach und Tumoren

In vielen Anwendungen macht es die Men-
ge der Daten und der vernetzten Parameter
sehr schwierig, das Verhalten der DNN zu
interpretieren. Selbst Schachspieler ha-
dern mit der mangelnden Transparenz
von Programmen, die DNN nutzen. Neu-
lich besiegte Google Alpha das beste her-
kömmliche Schach-Computerprogramm.
Aber niemand weiss so recht, wie das ge-
lang. Wenn es schon Schwierigkeiten beim
Schach gibt, wie steht es dann erst um me-
dizinische Hilfsprogramme zur Klassifika-
tion von Tumoren? Sind sie schon so ver-
ständlich und bewährt, dass man sich auf
die «Entscheidungen» der Computerhirne
verlassen möchte? Viele Forschende haben
da so ihre Zweifel - selbst wenn sie nicht
gleich von Alchemie sprechen würden.

Die Defense Advanced Research Projects
Agency des US-Verteidigungsministeri-
ums widmet sich bereits der Herausforde-
rung: Im Projekt «Explainable AI» werden
Modelle entwickelt, die auf DNN basieren,
aber dennoch für den Nutzer transparent
sind. Forschende an der Stanford University

wiederum haben neulich ein Programm
entwickelt, das neuronale Netze auf Fehler
untersuchen kann. Es eignet sich ausser-
dem dafür, die getroffenen Entscheidungen
besser zu verstehen. Das gelingt, indem die
Komplexität des Modells auf das Wesent-
liche reduziert wird.

Frossard und seine Gruppe verfolgen ein
anderes Konzept. Sie lassen empirisches
Vorwissen in ein DNN-gestütztes Modell
einfließen. Die Idee: Kombiniert man das
maschinelle Lernen mit konkreten Kennt-
nissen der Wirklichkeit, lässt sich womög-
lich ein Programm fertigen, das die Vortei-
le beider Seiten vereint - die Lernfähigkeit
von DNN mit der Interpretierbarkeit her-
kömmlicher Programme. Frossard: «Am
Ende hängt zwar alles von den Anwendun-
gen ab. Aber das beste System ist wahr-
scheinlich irgendwo in der Mitte.»

Sven Titz ist freier Wissenschaftsjournalist
in Berlin.

Modelle klauen und nachbauen

Ein spezielles Problem vielschichtiger
neuronaler Netze ist die Gefahr des Modell-
Diebstahls. Zwar werden die Programme
oft anhand von Daten trainiert, die geheim
sind. Durch einen Trick lassen sich die
Modelle aber nachbauen, ohne dass man
die Trainingsdaten kennt, erläutert Yannic
Kilcher von der ETH Zürich. Dazu stellt man
dem Modell «Fragen» (das sind zum Bei-
spiel Bilder im Fall eines Bilderkennungs-
algorithmus). Aus der Kombination mit
den Resultaten lässt sich - mithilfe eines
eigenen neuronalen Netzes - das Programm
nachbauen.

Das Problem besteht nun darin, dass
sich aus dem rekonstruierten Netzwerk
Informationen über die geheimen Trai-
ningsdaten ermitteln lassen. Wenn es sich
um Patientendaten handeln würde, wäre
das besonders heikel. Forschende wie
Kilcher haben aber bereits erste Versuche
unternommen, durch geschickte Verände-
rungen an den Programmen den Diebstahl
zu erschweren.