Projektbericht

Fabian Offert

Maschineninterpretation mit Interpretationsmaschinen

Explainable Artificial Intelligence als bildgebendes Verfahren und bildwissenschaftliches Problem

Maschinelles Sehen und künstliche Intelligenz sind historisch nicht voneinander zu trennen. Seymour Paperts *Summer Vision Project*[1] formulierte im Sommer 1966, zehn Jahre nach dem die Disziplin begründenden Dartmouth Workshop,[2] die Notwendigkeit, das menschliche Sehen auf der Basis von Gestaltprinzipien zu mechanisieren. Bereits 1958 hatte Frank Rosenblatt mit dem *Perceptron* die Möglichkeit, rudimentäres maschinelles Sehen mithilfe von künstlichen neuronalen Netzen zu realisieren, demonstriert.[3] Und es sind eben diese künstlichen neuronalen Netze, die heute die KI-Forschung bestimmen.

Künstliche neuronale Netze bauen auf dem Prinzip auf, Computer nicht zu programmieren, sondern zu trainieren, d. h. aus großen Datenmengen ‚lernen' zu lassen. Aus diesem Ansatz ergibt sich ein zentrales Problem: Während die Effizienz eines künstlichen neuronalen Netzes einfach bestimmbar ist, zum Beispiel indem man es mit von den Trainingsdaten unabhängigen Testdaten konfrontiert, ist es nur schwer bestimmbar, warum ein künstliches neuronales Netz effizient oder ineffizient arbeitet. Als rein probabilistisches Verfahren ist es einem künstlichen neuronalen Netz schlicht nicht anzusehen, wie es eine Aufgabe löst, selbst wenn es offensichtlich ist, dass es sie löst.

Was sind die Konsequenzen dieser Opazität künstlicher neuronaler Netze? 2015 spricht Frank Pasquale von der *Black Box Society,*[4] in der politisch und sozial relevante Entscheidungen von weitestgehend uneinsehbaren Maschinen getroffen werden. Das Aufkommen von *foundation models*[5] um 2020, also sehr großen, oft proprietären KI-Systemen, die nicht auf eine bestimmte Aufgabe spezialisiert sind, hat die bereits bei Pasquale thematisierten politischen und sozialen Risiken noch einmal besonders deutlich gemacht. *Foundation models* wie DALL-E 2 oder Stable Diffusion sind ganz offensichtlich in der Lage, ‚bedeutungsvolle' (hier verstanden als: nicht nur rein formal kohärente) Bilder zu produzieren, ohne dass es möglich wäre, die Prinzipien ihrer Entstehung auch nur im Ansatz nachzuvollziehen. Somit bleiben zum Beispiel dem Modell inhärente Vorurteile oft unsichtbar, und selbst wo sie sichtbar werden, können sie nicht im Nachhinein ausgeglichen werden.[6] Zahlreiche weitere politisch und sozial relevante Aspekte kommen nach und nach ans Licht, wie zum Beispiel das Problem der sogenannten *deep fakes*, also fiktiver Bilder real existierender Personen.

Die Informatik reflektiert diese Risiken nur selten in ihrer eigenen Forschung. Der so sich zunehmend verfestigende Eindruck einer sich beständig erweiternden Lücke bestimmt die zeitgenössische KI-Kritik. Von GeisteswissenschaftlerInnen im weitesten Sinne, so scheint es, wird vor allem die Aufarbeitung der sozialen und politischen Versäumnisse der Informatik erwartet. Übersehen wird dabei, dass es sich vor allem bei der Frage der Opazität um eine originär geisteswissenschaftliche Frage und, wie dieser kurze Artikel zeigen soll, auch um eine originär bildwissenschaftliche Frage handelt.

1 Seymour A. Papert: The Summer Vision Project, 1966, https://dspace.mit.edu/bitstream/handle/1721.1/6125/AIM-100.pdf (Stand 5/2023).
2 John McCarthy et al.: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf (Stand 5/2023).
3 Vgl. James Dobson: The Birth of Computer Vision, Minneapolis 2023.
4 Frank Pasquale: The Black Box Society. The Secret Algorithms That Control Money and Information, Cambridge/Mass. 2015.
5 Rishi Bommasani et al.: On the Opportunities and Risks of Foundation Models, arXiv preprint 2108.07258, 2021.
6 Vgl. Fabian Offert, Thao Phan: A Sign That Spells. DALL-E 2, Invisual Images and the Racial Politics of Feature Space, arXiv preprint 2211.06323, 2022.

https://doi.org/10.1515/9783111085692-010

Sie steht deshalb im Mittelpunkt eines an fünf Institutionen angesiedelten und von der Volkswagenstiftung geförderten Projektes, das in den kommenden drei Jahren die epistemischen, sozialen, politischen, und ästhetischen Aspekte von *explainable artificial intelligence* untersuchen möchte. *Explainable artificial intelligence*, gelegentlich auch bezeichnet als *interpretable machine learning*, meint jene Vielzahl von Verfahren, die in der Informatik entwickelt wurden und werden, um der Opazität von künstlichen neuronalen Netzen zu begegnen. Diese Verfahren, so die Arbeitshypothese des Projektes, können auf zweierlei Weise zur Kritik zeitgenössischer KI-Systeme in den Geisteswissenschaften beitragen. Zum einen können sie als technische Hilfsmittel einer bildwissenschaftlichen Aufarbeitung von spezifischen KI-Modellen dienen. Zum anderen müssen sie selbst als bildgebende Verfahren verstanden werden, die einer bildwissenschaftlichen Aufarbeitung bedürfen. Zentral ist dabei die Einsicht, dass beide Aspekte nur gemeinsam verhandelt und nur anhand spezifischer KI-Modelle diskutiert werden können.

Denn zum einen unterscheiden sich KI-Systeme untereinander so immens, dass keine allgemeingültigen Aussagen getroffen werden können. Zum anderen ist das Feld der *explainable artificial intelligence* in der Informatik geradezu gespalten: Wo es für einfache Prinzipien des maschinellen Lernens etablierte und gut funktionierende Ansätze gibt, die immer weiter verbessert werden,[7] versagen diese nicht erst bei *foundation models*, sondern bereits bei einfacheren bildverarbeitenden KI-Systemen. Als Beispiel soll hier eine Reihe von Experimenten dienen, die an den Universitäten Kaliforniens ihren Anfang fanden, in der Forschungsabteilung von OpenAI fortgesetzt wurden und mittlerweile zur Gründung einer dezidiert der Interpretierbarkeit von bildverarbeitenden KI-Systemen verschriebenen Firma namens Anthropic[8] geführt haben. Am Anfang dieser Experimente steht die Behauptung eines *empirical turns* in der *explainable artificial intelligence*:

> *„Just as the early microscope hinted at a new world of cells and microorganisms, visualizations of artificial neural networks have revealed tantalizing hints and glimpses of a rich inner world within our models [...]. This has led us to wonder: Is it possible that deep learning is at a similar, albeit more modest, transition point? [...] What if we were willing to spend thousands of hours tracing through every neuron and its connections? What kind of picture of neural networks would emerge?“*[9]

Die AutorInnen des zitierten Artikels sehen ihren Beitrag zur Forschung vor allem im Bereich *AI safety*, aber natürlich stellt eine solche Untersuchung Fragen in den Raum, die grundsätzliche, bildwissenschaftliche Fragen sind. Die zentrale dieser Fragen ist: Was können wir mithilfe von technischen Verfahren über die Weltwahrnehmung von KI-Systemen sagen, jenseits dessen, was sich im weitesten Sinne kybernetisch, also über den Vergleich von *input* und *output*, erfahren lässt?

In den oben erwähnten Experimenten der Gruppe um Chris Olah und Gabriel Goh kommt als zentrales technisches Verfahren die sogenannte *feature visualization* zum Einsatz. *Feature visualization* ist ein iteratives Verfahren, bei dem einzelne Neuronen eines neuronalen Netzes daraufhin überprüft werden, was ihre spezifische Rolle im Klassifizierungsprozess ist. Das Ergebnis sind nicht numerische Informationen, sondern Bilder, die zeigen, auf welche Bildbestandteile ein Neuron besonders reagiert.

7 Zum Beispiel SHAP und LIME, vgl. Christoph Molnar: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/ (Stand 5/2023).

8 Anthropic, https://www.anthropic.com (Stand 2/2023).

9 Nick Cammarata et al.: Thread: Circuits. In: Distill, 2020, https://distill.pub/2020/circuits/ (Stand 5/2023).

Solche Bilder enthalten Muster, Formen oder ganze Objekte, die auf eigentümliche Weise als Muster interpretiert werden. ↗ **Abb. 1** zeigt einen Vergleich von vier verschiedenen Architekturen im Hinblick auf Neuronen, die auf Kurven spezialisiert sind, und solche, die auf den Übergang von hohen und niedrigen Bildfrequenzen spezialisiert sind.

Was ist die epistemische Aussagekraft solcher Bilder? Auf den ersten Blick lassen sie eine hierarchische Struktur, in der Wahrnehmung reduktionistisch operationalisiert wird, vermuten.[10] Aus Grundprinzipien des Sehens, wie zum Beispiel der Fähigkeit, Kurven und bestimmte Texturen zu erkennen, entsteht in letzter Instanz die Fähigkeit, Objekte als solche wahrzunehmen. Die Frage, wie das analysierte Modell die ihm übertragene Aufgabe – in diesem Fall die Unterscheidung zwischen 1000 verschiedenen Bildkategorien[11] – löst, scheint damit beantwortbar: durch die gezielte Zerlegung eines Bildes in seine Bestandteile, beginnend mit einfachsten Mustern.

Eine genauere Untersuchung der eingesetzten Verfahren auf technischer Ebene zeigt jedoch eine diesen Bildern inhärente Paradoxie: Je deutlicher sie etwas zeigen, das vermeintlich technisch gesehen wird, desto weiter weg sind sie vom eigentlichen Prozess des technischen Sehens. Hier kommt die Frage der ‚Lesbarkeit' solcher Visualisierungen ins Spiel. Menschen erkennen Muster, wo auch nur der Ansatz eines Musters erkennbar ist (Pareidolie), und sehen in natürlich vorkommenden Phänomenen oft menschliche Züge. Im Gegensatz zu anderen Arten der Visualisierung, die Einsicht durch klar abgegrenzte Elemente versprechen, wird diese Fähigkeit bei der *feature visualization* geradezu eingefordert. *Feature visualizations* sind gerade eben lesbar – und dies auch nur durch gezielte technische Interventionen. Denn um sie lesbar zu machen, müssen im iterativen Optimierungsprozess, der dem Verfahren zugrunde liegt, regelmäßig bestimmte (hohe) Bildfrequenzen herausgefiltert werden. Wenn dies nicht geschieht, entsteht ein Bild, das nur noch als Rauschen wahrgenommen werden kann, da es ausschließlich aus hochfrequenten Informationen zusammengesetzt ist und damit ‚unlesbar' erscheint. Der Grund für diese Eigenheit liegt in einer technischen Präferenz fast aller neuronalen Architekturen: Texturen werden höher gewichtet als Formen. Konkret heißt das, dass zum Beispiel ein Bild eines Elefanten mit dem Fell einer Katze als Katze wahrgenommen wird, da die Katzentextur die Elefantenform überschreibt.[12] Die Lesbarkeit solcher Visualisierungen ist, mit anderen Worten, umgekehrt proportional zu ihrer Aussagekraft. Um Bilder zu produzieren, die überhaupt ‚etwas' zeigen, muss gerade die maschinelle Perspektive, die von nichtmenschlichen Wahrnehmungsprinzipien gekennzeichnet ist, aufgegeben werden.[13]

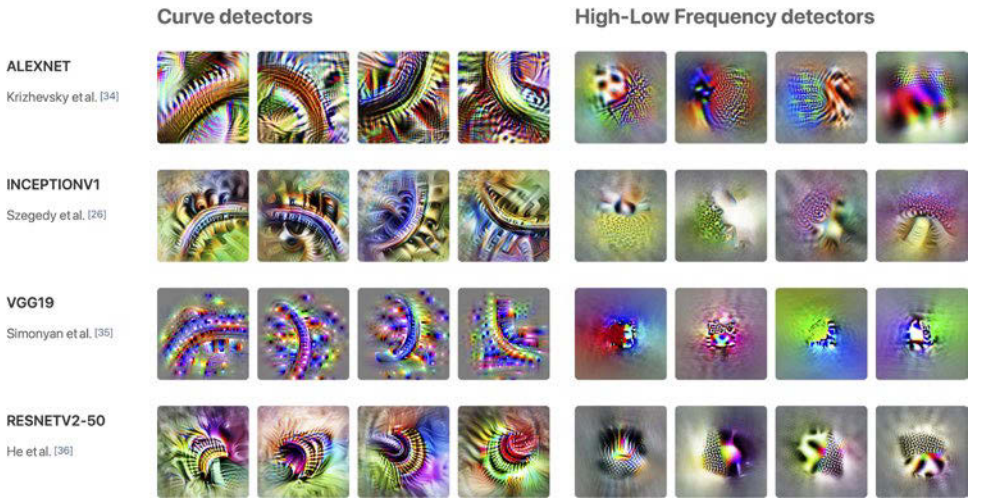
Es stellt sich damit die Frage, was solche technischen Bilder aus dem Bereich *explainable artificial intelligence* zu einer kritischen

10 Zur Frage der Überschneidung von technischen und biologischen Prinzipien der Wahrnehmung und ihren hypothetischen und tatsächlichen Hierarchien siehe David Marr: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, Cambridge/Mass. 1982. Zur Kritik dieser Frage im Kontext zeitgenössischer KI-Systeme vgl. Fabian Offert: Can We Read Neural Networks? Epistemic Implications of Two Historical Computer Science Papers. In: American Literature, Jg. 95, 2023, Heft 2, S. 423–428.

11 Die Trainingsdaten des Modells stammen aus dem ImageNet-Datensatz, dem wohl populärsten Trainingsdatensatz der KI-Forschung zwischen 2012 und 2020. ImageNet (in jener weithin verwendeten Variante, die im Rahmen der *ImageNet Large Scale Visual Recognition Challenge* 2012 erstellt wurde) versammelt 1,5 Millionen Bilder, die 1000 Kategorien zugeordnet sind.

12 Vgl. Robert Geirhos et al.: Shortcut Learning in Deep Neural Networks, arXiv preprint 2004.07780, 2020.

13 Eine ausführlichere Fassung dieser Argumentation findet sich in Fabian Offert, Peter Bell: Perceptual Bias and Technical Metapictures. Critical Machine Vision as a Humanities Challenge. In: AI & Society, 2021, Heft 36, S. 1133–1144.



1: Vergleichender Blick auf ein Subset von Filtern in verschiedenen Architekturen neuronaler Netzwerke, generiert mit dem feature-visualization-Verfahren. Aus: Cammarata et al. (2020, CC-BY 4.0). Die OpenAI-Microscope-Plattform (https://microscope.openai.com/models) ermöglicht es, verschiedene Architekturen interaktiv zu analysieren.

Aufarbeitung zeitgenössischer KI-Modelle beitragen können? Das Projekt versucht sich der Beantwortung dieser Frage in drei Schritten anzunähern. Der erste Schritt ist eine Bestandsaufnahme, in deren Rahmen alle bildlichen Verfahren aus dem Bereich *explainable artificial intelligence* zusammengetragen werden. Dazu gehört auch, ihre Einsatzgebiete zu erfassen, also in welchen Bereichen *explainable artificial intelligence* überhaupt eine Rolle spielt. Anhand dieser Taxonomie der Verfahren sollen in einem zweiten Schritt die konkreten (historischen und systematischen) bildwissenschaftlichen Aspekte der Verfahren herausgearbeitet werden: Wie werden Bilder produziert, was zeigen sie und wie zeigen sie es? Schließlich sollen in einem dritten und letzten Schritt ausgewählte Verfahren als interaktiver Werkzeugkasten mithilfe einer Web-Oberfläche zur Verfügung gestellt werden, eingebettet in eine umfassende Wissensdatenbank. *Explainable artificial intelligence* soll somit als Beitrag zu einer materialistischen Theorie und Praxis des maschinellen Lernens verstanden werden.