Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Abschlussbericht zum Master-Projekt

Vergleich verschiedener Verfahren, mit denen Entscheidungsprozesse von künstlicher Intelligenz transparent gemacht werden können

Fachbereich Informatik
Studiengang Master Informatik
Erstprüfer: Prof. Markus Ullmann
Zweitprüfer: Prof. Dr. Nico Hochgeschwender

eingereicht von:
Mario Beckel
Matr.-Nr. 9022094
Tenktererstr. 6
50679 Köln

Sankt Augustin, den 10.12.2019

Inhaltsverzeichnis

1 Einführung

Schon seit Jahrhunderten versuchen Menschen Maschinen zu bauen, die mit menschlichen Eigenschaften ausgestattet sind. Forscher sind schon lange von der Idee fasziniert Maschinen mit Künstlicher Intelligenz (KI) auszustatten. Was dabei unter künstlicher Intelligenz verstanden wird ist abhängig vom jeweiligen Forschungsstand und wie ähnlich Maschinen dem Menschen schon geworden sind. In den 1950er verstand man unter Künstlicher Intelligenz ein Automat, der schriftlich dividieren konnte. Heute wird dies als normal angesehen und würde nicht mehr unter dem Begriff KI fallen. [39, S. 23ff]

Seit Mitte des letzten Jahrhunderts existiert KI als anerkannte Wissenschaftsdisziplin. Das Forschungsgebiet beruht auf Kenntnissen aus ganz unterschiedlichen Gebieten wie Neurowissenschaften, Mathematik, Logik, Linguistik und Psychologie. In der zweiten Hälfte des letzten Jahrhunderts erlebte die Erforschung von KI verschiedene Phasen der Euphorie und der Ernüchterung. Nach einem interdisziplinären wissenschaftlichen Workshop im Jahr 1955 organisiert von den Wissenschaftlern Minsky, McCarthy und Shannon kamen große Erwartungen über einen Durchbruch von KI auf, die jedoch bald enttäuscht wurden. Darauf folgte viele Jahre der Ernüchterung. Theoretisches Wissen über die Umsetzung war zwar vorhanden, aber „die nötigen Prozesse, Werkstoffe und Energie“ fehlten noch. [39, S. 36] Die Umsetzung der Vorstellung einer maschinellen Intelligenz gestaltete sich schwieriger als erwartet.

In den letzten 15 Jahren wurden im Bereich der KI gravierende Fortschritte erzielt. Dies ist vor allem der wachsenden Rechenleistung von Computern zu verdanken. Durch den Einsatz von Grafikkarten (GPUs) und der Möglichkeit Rechenvorgänge parallel auszuführen, wurde die Verarbeitung von sehr großen Datenmengen möglich. Zudem erleichterte die wachsende Vernetzung der Computer und eine steigende Zahl von Internetnutzern, die Sammlung und Kategorisierung von großen Datenmengen.

Diese beide Entwicklungen ermöglichten Berechnungen mit größeren Modellen in komplexeren Architekturen. So konnten nun Algorithmen wie Neuronale Netze praktisch umgesetzt werden, deren Entwicklung bereits vor der Jahrtausendwende erfolgte. Zum Zeitpunkt der Entwicklung von Neuronalen Netzen, war die Realisierung aufgrund von begrenzten Rechenkapazitäten, noch nicht in dem Maße wie heute möglich.[2]

[2]Lecture 1: Introduction to Convolutional Neural Networks for Visual Recognition, Stanford University, 2017: https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv&index=1

Bereiche in denen KI bereits heute eingesetzt wird oder eine Einführung in nächster Zeit absehbar ist, sind zum Beispiel der Straßenverkehr (bei selbst fahrenden Autos), die Medizin (zur Einschätzung von Risikogruppen für bestimmte Krankheiten) und die Versicherungswirtschaft (zur Berechnung von Versicherungsprämien). Im Alltag begegnet uns KI, zum Beispiel in Smartphone Anwendungen, bei der Bearbeitung von Fotos, bei der Spracherkennung und in Online-Übersetzungstools.

Es gibt zahlreiche Einsatzgebiete, in denen neben Präzision noch die Erfüllung anderer Eigenschaften gefordert wird. Die Schaffung von Transparenz und Vertrauen beim Einsatz von komplexen Verfahren des maschinellen Lernens, ist eine Herausforderung. Modelle wie tiefe Neuronale Netze sind einer Black-Box vergleichbar, in der zwar die Eingabe- und die Ausgabewerte bekannt sind, aber der Prozess, wie die Vorhersage aus den Eingangsparametern erzeugt wurde, im Dunkeln verbleibt. Selbst für Experten ist der komplexe Prozess der Entscheidungsfindung bei diesen Modellen schwer nachvollziehbar. Hier wäre es wünschenswert, dass dieser Prozess transparenter gestaltet wird.

1.1 Motivation

Je stärker künstlich intelligente Entscheidungssysteme in den Alltag der Menschen Einzug halten, desto stärker treten Ängste und Zweifel gegenüber solchen Systemen hervor. Dies ist nachvollziehbar, da es für Laien nahezu unmöglich ist, die Chancen und Risiken von KI-Systemen abzuschätzen.

Als Reaktion auf drohende Gefahren bei dem Einsatz von Künstlicher Intelligenz in Entscheidungssystemen wurde unter anderem die Datenschutz-Grundverordnung eingeführt. Aus dieser kann ein Recht auf Erklärbarkeit abgeleitet werden, damit die von einer Entscheidung Betroffenen nachvollziehen können, wie eine Entscheidung zustande gekommen ist.

Aus technischer Sicht ist Erklärbarkeit bis heute kein Qualitätskriterium für die Beurteilung der Qualität eines Verfahrens des Maschinellen Lernens. Dennoch erläutert [9], dass Erklärbarkeit und Transparenz in mehrfacher Hinsicht auch für die Funktionalität von KI-Modellen vorteilhaft sein kann. Die Autorin sieht folgende Vorteile bei transparenter KI:

1. **Das Modell verbessern:** Sind die Grundlagen der Entscheidungen oder Vorhersagen bekannt, dann kann die Sinnhaftigkeit der beim Training des Modells verwendeten Regeln analysiert werden. Die Transparenz von KI-Systemen schützt davor, falsche Schlüsse zu ziehen. Wie Ribeiro in einem Experiment, in dem eine Bildklassifizierung durchgeführt werden sollte, zeigte, kann der Hintergrundbereich eines Bildes, der ausschlaggebende Impuls für die Entscheidung des Modells sein und nicht das eigentliche Objekt. So wurde die Entscheidung, ob im Bild ein Husky oder einen Wolf zu sehen ist, aufgrund des Schnees im Hintergrund, zugunsten des Huskys gefällt. Die Ursache dafür

war, dass im Trainingsdatensatz nahezu alle Bilder, in denen ein Husky zu sehen war, in einer Schneelandschaft aufgenommen wurden. [26, S. 8f]

2. **Vertrauen und Akzeptanz ermöglichen:** Menschen vertrauen der Vorhersage eines Modells eher, wenn sie den kausalen Zusammenhang, der zu der Entscheidung geführt hat, verstehen können. Dies ist besonders in Bereichen wichtig, in denen eine Fehlentscheidung zu gravierenden Schäden führen kann, also wenn die Entscheidung Einfluss auf das Leben oder die Gesundheit eines Menschen hat oder bedeutende finanzielle Folgen haben kann. [18, Kap. 2.1]

3. **Vorurteile und Fehler im Modell beseitigen:** Modelle, die mit Datensätzen trainiert wurden, die versteckte Vorurteile enthalten, können zu einer selbst erfüllenden Prophezeiung oder zur Stärkung von vorhandenen Diskriminierungen führen. Es können kausale Zusammenhänge suggeriert werden, wo gar keine sind. Cathy O'Neil hat in ihrem Buch viele Beispiele aufgeführt, die zeigen, dass durch einen nicht reflektierten Einsatz von KI, großer Schaden angerichtet werden kann.[1]

Da der Einsatz von Künstlicher Intelligenz in vielen Fällen vorteilhaft sein kann, sollten Wege gefunden werden, diesen so zu gestalten, dass die Menschen der Technik vertrauen können. Dazu ist es notwendig, dass die Betroffenen nachvollziehen können, wie eine Entscheidung zustande gekommen ist. Welche Methoden existieren um den Entscheidungsprozess transparenter zu machen, welche Eigenschaften diese Methoden haben und wie sie sich unterscheiden, wird im weiteren Verlauf der Arbeit untersucht.

1.2 Zielsetzung

In den letzten Jahren sind zahlreiche Ansätze und Methoden zur Erzeugung von Transparenz in Black-Box Modellen veröffentlicht worden. Ziel dieses Masterprojekts ist es, diese Ansätze und Methoden zu ermitteln, sie zu analysieren und zu vergleichen.

Unter „Ansatz" wird eine übergeordnetes Vorgehen verstanden, das von unterschiedlichen Methoden eingesetzt wird. Für jeden vorgestellten Ansatz werden in dieser Arbeit verschiedene konkrete Methoden mit ihrer Funktionsweise und den jeweiligen Eigenschaften beschrieben. Es werden Kriterien vorgestellt, mit den die verschiedenen Methoden bewertet und verglichen werden können.

Ein wichtiges Unterscheidungskriterium sind die geeigneten Einsatzgebiete der jeweiligen Methoden. Wichtige Einsatzbereiche für Modelle der künstlichen Intelligenz sind die Text-, Bild- und die Videoanalyse. Da die anschließende Masterarbeit ihren thematischen Schwerpunkt in der Bildverarbeitung haben wird, sind für dieses

[1] Cathy O'Neil: Weapons of Math Destruction, 2016

Projekt die Methoden von besonderer Bedeutung, die in diesem Bereich eingesetzt werden können. Die Masterarbeit soll auf den im Masterprojekt erworbenen Kenntnissen aufbauen.

1.3 Aufbau der Arbeit

Zunächst wird im Kapitel 2 erläutert was unter „Erklärbarkeit“ verstanden wird und wie der Begriff von Transparenz abgegrenzt werden kann. Nach der Definition der beiden Begriffe werden die Vorteile, die Erzeugung von Erklärbarkeit bei KI-Modellen beschrieben. Im Anschluss wird auf die verschiedenen Dimensionen eines KI-Modells eingegangen, um zu erläutern, in welchen Bereichen Transparenz hergestellt werden kann.

Im Fokus der weiteren Betrachtung stehen die Methoden, die zur Erzeugung von Erklärbarkeit herangezogen werden können. In diesem Zusammenhang werden verschiedene Klassifizierungsmöglichkeiten für KI-Methoden beschrieben. Zum Schluss des Kapitels werden Kriterien erläutert, mit den unterschiedliche Erklärungsmethoden bewertet werden können.

Das folgende Kapitel 3 stellt unterschiedliche Erklärungsansätze vor, mit denen nachvollzogen werden kann, wie ein KI-Modell zu einer Entscheidung gekommen ist. Die Ansätze können als grundlegende Ideen verstanden werden, wie Erklärbarkeit hergestellt werden kann. Die Autoren von Methoden setzen meist einen oder mehrere dieser Ansätze ein, um Erklärbarkeit herzustellen. In diesem Kapitel werden einzelne Ansätze beschrieben. Es wird erörtert, welche Methoden nach welchem Ansatz arbeitet und ob einzelne Methoden verschiedene Ansätze miteinander kombinieren. Die einzelnen Methoden werden in diesem Kapitel kurz vorgestellt, eine genauere Beschreibung findet sich im Anhang dieses Berichts.

In Kapitel 4 werden die zuvor erläuterten Ansätze miteinander verglichen und die Vor- und Nachteile der einzelnen Ansätze herausgearbeitet.

Das letzte Kapitel 5 geht auf den zukünftigen Entwicklungsbedarf der zur Erklärung eingesetzten Methoden ein und beschreibt den potentiellen Nutzen von Erklärbarkeits-Methoden. In diesem Kapitel wird auch eine abschließende Bewertung des Master-Projekts vorgenommen.

2 Grundlagen

Eine allgemein anerkannte einheitliche Definition von Interpretierbarkeit und Erklärbarkeit existiert bis heute nicht. Auch wenn es verschiedene Versuche der Definition gibt, konnte sich bisher keine Definition durchsetzen. Exemplarisch seien hier zwei Definitionen aufgeführt. Montavon et al. definiert in [20, S. 2] die beiden Begriffe. Interpretierbarkeit definiert er folgendermaßen:

> Eine **Interpretation** ist die Abbildung eines abstrakten Konzepts (zum Beispiel einer vorhergesagten Klasse) auf einen Bereich, den der Mensch verstehen kann. [20, S. 2]

Zum Beispiel können Bilder (die als Pixeln gesehen werden können) und Texte interpretiert werden. Ein Mensch kann Bilder betrachten und Texte lesen. Nicht interpretierbar sind zum Beispiel abstrakte Vektorräume oder Sequenzen von unbekannten Wörtern und Symbolen.

Demgegenüber grenzt der Autor den Begriff der Erklärbarkeit ab:

> Eine **Erklärung** umfasst die gesamten Merkmale des zu interpretierenden Bereichs, die . . . dazu beigetragen haben, eine Entscheidung zu treffen (zum Beispiel eine Klassifikation oder eine Regression) [20, S. 2]

Die Merkmale die für eine Erklärung ausschlaggebend sind, können hinsichtlich ihrer Relevanz, die sie für die Erzeugung der Vorhersage haben, bewertet werden. Die anschließende Erklärung kann über eine Heatmap erfolgen. In der Heatmap werden die Bereiche des Bildes hervorgehoben, die den größten Einfluss auf die Klassifikation gehabt haben.

Im Bereich der Klassifizierung von Bildern, wird jedes einzelne Pixel als Merkmal betrachtet. Ein Bild mit einer Größe von 100 mal 100 Pixeln ist also mit 10.000 Merkmale ausgestattet.[1]

Nicht alle Autoren die sich dem Thema Erklärbarkeit von Künstlicher Intelligenz widmen, folgen diesen Definitionen. Viele Autoren verwenden die beiden Begriffe synonym. Da der Unterschied der beiden Definitionen subtil ist und die Unterscheidung der beiden Begriffe für diese Arbeit nicht ausschlaggebend ist, werden die Begriffe hier synonym verwendet.

Die Erklärbarkeit von KI kann auch in Bezug auf das zu untersuchende Objekt unterschieden werden. Diese Arbeit befasst sich schwerpunktmäßig mit der **Erklärung des Entscheidungsprozesses**. Aber es werden auch Methoden vorgestellt, die das **Modell** erklären.

[1] Andriy Burkov: Machine Learning kompakt, 2019, S. 95

2.1 Nutzen von Erklärbarkeit

Nachfolgend werden die Gründe dafür erläutert, warum es vorteilhaft ist, auf Algorithmen beruhende Entscheidungssysteme transparent zu gestalten. Die Ausführungen beruhen auf einem Vortrag den Herr Samek am 17. September 2019 im Bundesamt für Sicherheit in der Informationstechnik (BSI) gehalten hat.

1. **Korrektheit**
   Es soll sichergestellt werden, dass die Klassifizierungsmethode so wie erwartet funktioniert. Wenn das Modell falsche Entscheidungen trifft kann dies gefährlich und teuer sein, zum Beispiel können autonom fahrende Autos einen Unfall verursachen, wenn sie Hindernisse auf der Straße nicht erkennen oder falsch interpretieren. Fehlerhafte Klassifizierungsentscheidungen von KI bei medizinischen Diagnosen, können zu falschen Behandlungen führen.

2. **Vertrauen**
   Menschen können einem maschinellen Verfahren mehr Vertrauen entgegenbringen, wenn für sie nachvollziehbar ist, wie ein Entscheidungsprozess abläuft und welche Kriterien eine Entscheidung beeinflussen.

3. **Neue Einsichten**
   Menschen können ihr Wissen durch einen transparenten Entscheidungsprozess eines Modells des Maschinellen Lernens erweitern. So hat zum Beispiel das auf KI beruhende Computerprogramm AlphaGo im März 2016 den weltweit besten Profispieler besiegt. Entscheidend zum Sieg beigetragen haben Spielzüge, die der Fachwelt bisher noch unbekannt waren.[2] Eine transparente KI kann zu neuen wissenschaftlichen Einsichten in Biologie, Physik, Chemie und anderen Bereichen führen.

4. **Verbesserung des Modells**
   Wenn das Vorgehen eines KI-Modells für Menschen nachvollziehbar ist, können Bereiche des Modells ermittelt werden, in denen das Modell noch nicht optimal arbeitet. Zusätzlich kann Expertenwissen aus der entsprechenden Domäne an der passenden Stelle in das Modell implementiert werden, um das Modell zu optimieren.

   Bei der Erzeugung von Transparenz in Neuronalen Netzen können versteckte Merkmale zum Vorschein kommen, die in den Eingangsdaten nicht vorhanden sind und die erst durch das Zusammenwirken von verschiedenen Eingangsdaten erzeugt worden sind.[43, S. 1]

5. **Einhalten von Gesetzen**
   Nach der Datenschutz Grundverordnung haben die Menschen in der EU ein Recht auf Erklärung. Transparente Künstliche Intelligenz kann helfen, dieses Recht umzusetzen. Ein Beispiel dafür sind Entscheidungen bei der Kreditver-

[2] https://de.wikipedia.org/wiki/AlphaGo

gabe. In diesem Fall hilft Transparenz zu verstehen, warum das Modell sich für oder gegen die Kreditvergabe an eine bestimmte Person ausgesprochen hat.

2.2 Verschiedene Dimensionen der Erklärungen

Im Allgemeinen kann Interpretierbarkeit auf den gesamten AI-Entwicklungsprozess angewendet werden. Besonders kann sie vor der Erstellung des Modells, während der Ausführung des Modells und zur Erklärung des Resultats eingesetzt werden. Die Abbildung 2.1 veranschaulicht die drei Dimensionen. In der ersten Dimension

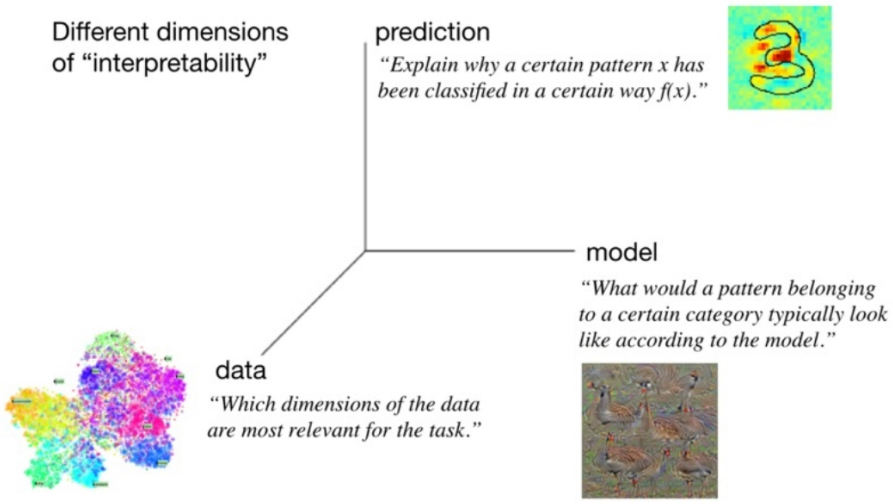


Abbildung 2.1: Dimensionen der Erklärbarkeit [29, S.7]

der Analyse liegt der Schwerpunkt darauf, die dem Modell zugrunde liegenden Daten zu verstehen. Es ist von Interesse, zu begreifen, welche Daten für die Vorhersagen die höchste Relevanz haben. In der zweiten Dimension steht die Transparenz des Entscheidungsprozesses des Modells im Vordergrund. Es geht darum zu verstehen, wie die Vorhersage zustande gekommen ist. Das Modell steht im Fokus der dritten Dimension. Hier ist die Aufgabe zu verstehen, wie das Modell funktioniert.

Für alle drei Phasen gibt es Methoden, mit denen eine Transparenz hergestellt werden kann. Der Schwerpunkt dieser Arbeit liegt darauf Methoden vorzustellen, die erklären können, wie eine Entscheidung zustande gekommen ist. Es werden auch einige Methoden in Abschnitt 3.2.5 präsentiert, die das Modell erklären.

Unterschiedliche Empfänger haben verschiedene Erwartungen an eine Erklärung über die Entscheidungsfindung eines Black-Box Modells. KI-Experten benötigen mechanistische Erklärungen. Sie wollen wissen, wie die verschiedenen Layer eines tiefen Neuronalen Netzes auf die Eingabedaten reagieren. Hier steht die Beziehung der Eingangsdaten zu den Layern des Netzwerkes im Vordergrund.

Nicht-Experten benötigen eher funktionale Erklärungen. Für sie ist entscheidend zu wissen, wie das Black-Box Modell die Eingangsdaten zur Vorhersage verwendet. [13]

2.3 Klassifizierung von Interpretations-Methoden

Es existiert eine Vielzahl von Methoden, die sich zum Ziel gesetzt haben, die Prozesse des Maschinellen Lernens erklärbar zu machen und fast täglich erscheinen neue wissenschaftliche Arbeiten, die sich mit dem Thema befassen. Einige Autoren [18, 30] haben versucht eine Taxonomie der Erklärbarkeitsansätze zu erstellen, doch nicht immer sind alle Methoden eindeutig einem Ansatz zuzuordnen. Hier werden zunächst drei Kriterien erläutert, mit denen eine Erklärungsmethode grob klassifiziert werden kann. Die geschilderten Gegensatzpaare werden in [18] erläutert.

Intrinsisch oder Post-hoc

Die Erklärbarkeit eines Modells kann auf zwei verschiedene Arten hergestellt werden. Das Vorgehensmodell von intrinsischen Modellen ist auf Grund ihrer einfachen Struktur leicht nachvollziehbar. Zu dieser Kategorie gehören zum Beispiel flache Entscheidungsbäume und einfache lineare Modelle.

Post-hoc Modelle sind in der Regel komplexer. Nach der Trainingsphase (post-hoc) werden sie durch die Verwendung von zusätzlichen Analysemethoden erklärbar. Zu dieser Kategorie gehören Methoden wie Surrogat Modelle, LIME und Layerwise Relevance Propagation (LRP). Post-hoc-Methoden können auch Erklärungen für intrinsisch interpretierbare Modelle erzeugen.

Modell-spezifisch oder Modell-agnostisch

Modell-spezifische Erklärungsansätze können nur Erklärungen für eine bestimmte Modellklasse erzeugen. Da intrinsisch interpretierbare Methoden immer nur auf ein Modell ausgerichtet sind, sind sie Modell-spezifisch. Darüber hinaus gibt es Methoden, die nur für eine bestimmte Art von Modellen anwendbar sind.

Modell-agnostische Methoden können jedes Modell des Maschinellen Lernens erklären. Diese Methoden verwenden die Zuordnungen zwischen Eingabe- und Ausgabewerten um Erklärungen zu erzeugen. Dazu verwenden sie keine internen Informationen, wie Gewichte und strukturelle Eigenschaften, des analysierten Modells.

Obwohl modellunabhängige Interpretationsmethoden praktisch sind, verwenden sie oft Ersatzmodelle oder erzeugen anderweitig Näherungswerte. Dies kann die Genauigkeit der erzeugten Erklärungen beeinträchtigen. Modellspezifische Interpretationstechniken gründen ihre Erklärungen oft direkt auf dem zu interpretierenden Modell und können daher genauer sein. [8, S. 14]

Lokale oder globale Erklärungen
Mit diesem Kriterium wird unterschieden, ob die Interpretationsmethode das gesamte Modellverhalten erklärt oder nur eine einzelne Vorhersage. Lokale Methoden erläutern die Vorhersage für einen einzelnen Datenpunkt. Im Gegensatz dazu betrachten globale Modelle das vollständige Modell und versuchen den Entscheidungsprozess als Ganzes zu erfassen.

Globale Erklärungen können in einigen Fällen ungenau sein. Lokale Erklärungen haben den Vorteil, dass für kleine Bereiche des Modells eine Annäherung über eine lineare oder monotone Funktion möglich ist und somit eine höhere Genauigkeit erreicht werden kann. Oft liefert eine Erklärung mittels einer Kombination von lokalen und globalen Interpretationstechniken die besten Ergebnisse. [8, S.13]

2.4 Eigenschaften von Erklärungen

Vorhersagen von Modellen des Maschinellen Lernens können mit Hilfe von Algorithmen erklärbar gemacht werden. Um eine für den Menschen verständliche Erklärung zu generieren, wird die Beziehung zwischen den Eingabe-Merkmalen und der Vorhersage des Modells verwendet. Die Autoren [27, S.162] und [18, Kapitel 2.5] haben zur Bewertung der Qualität von Erklärungsmodellen verschiedene Kriterien aufgestellt, von denen die Wichtigsten im Folgenden näher erläutert werden. Ein Bewertungsmodell mit dem der Grad der Einhaltung dieser Kriterien beurteilt werden könnte, existiert bisher noch nicht. [18, Kapitel 2.5]

Exaktheit Ein Modell arbeitet exakt, wenn es auch für bisher unbekannte Instanzen gute Vorhersagen erzeugt. Arbeit ein Modell zum Beispiel regelbasiert und die Regeln sind so allgemein formuliert, dass sie auch für bisher unbekannte Instanzen richtige Vorsagen erzeugen, dann wird die Vorgehensweise des Modells als exakt beschrieben. [27, S.162]

Die Exaktheit ist von Bedeutung, wenn das Original-Modell durch ein anderes, interpretierbares Modell, ersetzt werden soll. Die Ersatzmodelle werden auch als Surrogat-Modelle bezeichnet. Dies ist ein Ansatz um Erklärbarkeit zu erzeugen, der in Abschnitt 3.2.1 näher beschrieben wird.

Wiedergabetreue ist ein Maß für die Annäherung der Erklärung an die Klassifikationsentscheidung des Black-Box Modells. Wiedergabetreue und Exaktheit liegen nah beieinander. Wenn ein Black-Box Modell exakte Vorhersagen trifft und die Erklärung von hoher Wiedergabetreue ist, dann ist auch die Erklärung exakt.

Man kann zwischen lokaler und globaler Wiedergabetreue unterscheiden. Bei einer lokalen Wiedergabetreue ist die Annäherung nur für einen Teilbereich passend, die globale Wiedertreue umfasst dagegen den gesamten Kontext des Modells.

Konsistenz ist eine Information darüber, ob ähnliche Erklärungen von unterschiedlichen Modellen erzeugt werden, die auf dieselbe Aufgabe trainiert wurden. Konsistenz ist gegeben, wenn die Erklärungen gleichartig sind.

Stabilität Im Gegensatz zur Konsistenz, bei der mehrere Black-Box-Modelle miteinander verglichen werden, werden bei der Stabilität verschiedene Erklärungen eines Modells miteinander verglichen. Die Stabilität eines Verfahrens ist hoch, wenn für ähnliche Instanzen ähnliche Erklärungen erzeugt werden, das heißt, eine geringe Änderung der Eingabewerte sollte nicht zu einer wesentlichen Änderung der Erklärung führen.

Verständlichkeit informiert darüber, wie gut die erzeugten Erklärungen für einen Menschen verständlich sind. Diese Eigenschaft ist wichtig, da sie das Ziel des Erklärungsmodells ist, sie ist aber oft schwer zu messen, weil sie stark vom Wissensstand des Empfängers abhängt.

Gewissheit Diese Eigenschaft liefert Informationen darüber, mit welcher Wahrscheinlichkeit die erzeugten Erklärungen einer Methode korrekt sind.

Grad der Wichtigkeit Wird durch die erzeugte Erklärung erkennbar, welche Eingabe-Merkmale oder welche Aspekte der Erklärung wichtig sind? Ist zum Beispiel erkennbar welche Variablen und welche Regeln für eine Erklärung am wichtigsten waren?

Leider findet man zu diesen Kriterien in den wissenschaftlichen Arbeiten, die die einzelnen Methoden beschreiben, wenig Informationen. Dies mag zum einen darin begründet sein, dass sich die wissenschaftliche Gemeinschaft bisher nicht auf einen allgemeingültigen Kriterienkatalog zur Beschreibung von Methoden geeinigt hat, zum anderen sind die Wissenschaftler oft daran interessiert, ihre Methode möglichst positiv darzustellen. Eine Einschätzung darüber, bis zu welchem Grad jede Methode diese Eigenschaften erfüllt, würde eine genauere Analyse der einzelnen Methoden erfordern.

3 Erklärbarkeit erzeugen

Im ersten Teil dieses Kapitels werden Algorithmen vorgestellt, die die Erklärbarkeit eines Modells erzeugen können. Daran anschließend erfolgt im nächsten Abschnitt

eine Darstellung der verschiedene Erklärungsansätze, um entweder die Entscheidung eines Modells zu erklären oder um das Modell selbst zu erklären.

3.1 Interpretierbare Modelle

Der einfachste Möglichkeit Interpretierbarkeit herzustellen, besteht darin, Methoden zu verwenden, die in sich schon so verständlich sind, so dass ein Mensch den Vorhersageprozess leicht nachvollziehen kann. Einige intrinsische erklärbare Methoden werden in diesem Abschnitt vorgestellt. Mehrere der später vorgestellten komplexeren Methoden setzen eine der hier vorgestellten interpretierbaren Modelle ein. So kommen zum Beispiel Entscheidungsregeln und -bäume in einigen Methoden zum Einsatz, um Erklärbarkeit herzustellen.

Für die Erzeugung von Transparenz bei einem Black-Box Modell gibt es unterschiedliche Vorgehensweisen. Die Wahl des richtigen Ansatzes ist ausschlaggebend für den Informationsgehalt, den ein Nutzer daraus ziehen kann. Ein guter und vollständiger Erklärungsansatz wird der Komplexität des zu erklärenden Modells gerecht und kann exakte Vorhersagen erzeugen. Oft muss bei der Wahl des richtigen Ansatzes einen Kompromiss zwischen dem Grad der Interpretierbarkeit und dem Grad der Vollständigkeit eingegangen werden. [13]

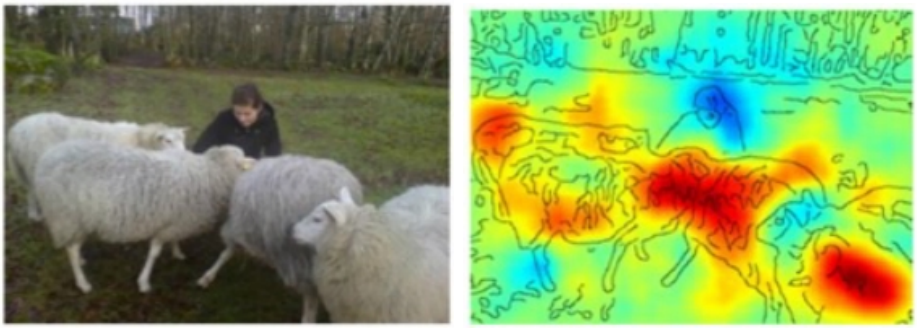
Relevanzwerte

Ein häufig genutztes Vorgehen um ein Black-Box Modell transparent zu gestalten sind Relevanzwerte. Bei diesem Verfahren wird der Beitrag, den jedes Merkmal zur Vorhersage geleistet hat, ermittelt. Die einzelnen Beiträgen können abhängig von der eingesetzten Methode visuell in Form von Saliency Heatmaps oder numerisch über Shapley Values dargestellt werden. [13]

In der Bilderkennung werden Saliency Heatmaps eingesetzt, um die für die Vorhersagen wichtigen Bereiche eines Bildes visuell hervorzuheben. Eine **Heatmap** ist ein Bild (oder eine graphische Darstellung) in der wichtige Bereiche farblich hervorgehoben werden (siehe Abbildung 3.1). Im Zusammenhang mit der Klassifizierung von Bildern, werden in einer Heatmap Pixel rot dargestellt, die für eine Vorhersage relevant sind. Blaue Pixel stehen im Widerspruch zur Vorhersage und graue Pixel haben keinen Einfluss. [19, S. 254]

Entscheidungsregeln

Entscheidungsregeln sind meist nachvollziehbar und können von anderen Methoden zur Erklärung herangezogen werden. Jede Entscheidungsregel hat im Allgemeinen die Form einer „wenn-dann-Bedingung“. Die Bedingung ist eine einfache Funktion, die aus den Eingangsmerkmalen die Vorhersage erstellt.



(a) Das zu klassifizierende Bild

(b) Die Heatmap des linken Bildes (die Bereiche, die für eine Klassifizierung „Schaf“ sprechen sind rot markiert)

Abbildung 3.1: Darstelltung einer Heatmap zur Klassifizierung. [29, S.12]

Es existieren Methoden, die über einen Satz von Regeln die Funktionsweise tiefer Neuronaler Netze beschreiben. Meistens ist dieser Ansatz für die Anwendung auf komplexe moderne Neuronale Netze nicht geeignet. Ribeiro et. al haben mit Anchors eine Methode vorgeschlagen, die über lokale Regeln Instanzvorhersagen erstellen kann. [13]

Entscheidungsbäume

Entscheidungsbäume haben eine gewisse Ähnlichkeit zu Entscheidungsregeln. Im Gegensatz zu den Regeln haben die Entscheidungsbäume die Struktur eines Graphen. In den Knoten finden sich die Bedingungsprüfungen für die Eingabemerkmale und in den Blättern sind die Ergebnisse des Modells zu finden. Anders als bei den Entscheidungsregeln gibt es hier immer nur einen Pfad vom Wurzelknoten zu einem Blatt. Über die DeepRED-Methode können tiefe Neuronale Netze mit Entscheidungsbäumen approximiert werden.[13]

Es existieren verschiedene Techniken mit deren Hilfe Entscheidungsbäume in Entscheidungsregeln transformiert werden können. [10, S.17]

Abhängigkeitsdiagramme (Dependency Plots)

Abhängigkeitsdiagramme stellen die Beziehung zwischen Eingabewerten und den erzeugten Ausgabewerten dar. Mit anderen Worten zeigen sie die Änderung der Vorhersage bei modifizierten Eingangswerten. Beispielsweise basieren die Methoden Partial Dependence Plot (PDP) und Individual Conditional Expectation (ICE) auf diesem Ansatz. [13]

Mit PDPs ist es möglich die Beziehung von ein oder zwei Eingangsvariablen zum Ergebnis des Black-Box Modells graphisch darzustellen. [8, S. 26] Dabei kann die Art der Beziehung zwischen Eingabe- und Ausgabewert ermittelt werden, ob sie linear, monoton oder komplex ist. [18, S. 118]

ICEs funktionieren ähnlich wie PDPs. Während bei PDPs der Bezugspunkt für die Beziehung zwischen Eingabe- und Ausgabe-Werte der Mittelwert über alle Instanzen ist, wird bei der ICE-Methode die Beziehung für jede Instanz separat dargestellt. Im Gegensatz zum PDP, bei dem ein Linienverlauf zu sehen ist, werden bei ICE viele Verläufe dargestellt. [18, S. 119]

Architektur-spezifische Visualisierungen

Um die innere Funktionsweise eines Modells transparent zu machen können Visualisierungen erzeugt werden, die der inneren Struktur des Black-Box Modells angepasst sind. Methoden, die diesen Ansatz verwenden sind LSTMViz und GAN Lab, auf die in dieser Arbeit jedoch nicht weiter eingegangen wird. [13]

Kontrafaktische Erklärungen

Dieser Ansatz untersucht, welche Eingangswerte in welcher Weise geändert werden müssen, damit ein gewünschtes Vorhersage-Ergebnis erzielt wird. Ein Beispiel: Durch ein Black-Box Modell wurde die Entscheidung aufgrund von den eingegebenen Kundendaten getroffen, dass ein Kunde nicht kreditwürdig ist. Mit dem Ansatz der Kontrafaktischen Erklärungen wird nun geprüft, welche Kundendaten wie geändert werden müssten, damit der Kunde doch noch einen Kredit der Bank bekommen kann. Zum Beispiel könnte sich der Besitz von weniger Kreditkarten positiv auf die Entscheidung auswirken. [13]

3.2 Verschiedene Erklärungsansätze

Aus der Vielzahl der Methoden, die sich mit der Erzeugung von Erklärbarkeit für KI-Modelle befassen, lassen sich verschiedene Ansätze erkennen, die die unterschiedlichen Methoden einsetzen. Einige Methoden haben eine ähnliche Vorgehensweise oder unterscheiden sich nur in einigen Details. Um eine bessere Übersicht zu erlangen, können Methoden, die einen ähnlichen Ansatz verfolgen in einer Klasse zusammengefasst werden. Bisher gibt es aber noch keine allgemeingültige Klassifizierung der Methoden und auch nur wenige wissenschaftliche Arbeiten, die sich dieser Aufgabe gewidmet haben.

Manche Methoden lassen sich nicht eindeutig einem Ansatz zuordnen, weil sie verschiedene Verfahren miteinander kombinieren. In dieser Arbeit habe ich weitgehend die Klassifizierung von Samek et al. übernommen, wie sie in [30] erläutert wird.

Eine erste grobe Unterscheidung zwischen verschiedenen Erklärungsansätzen kann dahingehend vorgenommen werden, ob eine Methode das Verhalten einer Black-Box erklärt oder das Black-Box Modell an sich.

Diese Arbeit widmet sich hauptsächlich Methoden, die das Verhalten eines KI-Modells erklären. Um Erklärbarkeit herzustellen können entweder Surrogat-Modelle

(3.2.1) oder Gradienten-basierte (3.2.2.1), Störungs-basierte (3.2.2.2), Optimierungsbasierte (3.2.3) oder Ausbreitungs-basierte Modelle (3.2.4) eingesetzt werden. Eine Übersicht über die in dieser Arbeit beschriebenen Erklärungsansätze und die Methoden, die auf diesen Ansätzen beruhen, bietet die Abbildung 3.2. In dieser Darstellung sind im unteren Bereich auch die Methoden erwähnt, die sich der Erklärung des Modells zum Ziel (3.2.5) gesetzt haben.

3.2.1 Erklärung über Surrogate-Modelle

Lineare Modelle und flache Entscheidungsbäume sind in sich interpretierbar. Hier werden keine weiteren Methoden gebraucht, um die Vorhersagen zu erklären. Anders sieht es bei komplexen Modellen wie tiefen Neuronalen Netzen aus, bei denen Entscheidungen über mehrere Schichten und nicht-lineare Funktionen zustande kommen. Aufgrund der komplexen Struktur dieser Modelle, ist eine Transparenz des Entscheidungsprozesses oft nicht gegeben. [30, S. 12]

Eine Möglichkeit Interpretierbarkeit herzustellen ist der Einsatz von Surrogate Modellen. Die Idee ist, das komplexe Ursprungsmodell durch ein einfacheres, interpretierbares Ersatzmodell nachzuahmen. [31, S. 12] Dazu muss ein Ersatzmodell gefunden werden, dass die Vorhersage des Ausgangsmodells so genau wie möglich abbildet und das leicht interpretierbar ist. Um das Surrogat Modell zu trainieren, sind keine Informationen über die innere Struktur des abzubildenden Black-Box Modells notwendig. Das Surrogat Modell wird mit Hilfe der Daten und der Vorhersagefunktion des Ausgangsmodells erzeugt. Die Unabhängigkeit von der inneren Struktur des Modells hat den Vorteil, dass das Surrogat Modell weiterhin verwendet werden kann, auch wenn das Black Box Modell ausgetauscht wird, solange die Eingabewerte und die Vorhersage nicht geändert werden. [18, Kap. 5.6]

Surrogate Modelle können durch lineare Regression oder durch die Erstellung von Entscheidungsbäumen über die ursprünglichen Eingangsdaten und der Vorhersagefunktion erzeugt werden. Mit den Ersatzmodellen können sowohl globale, als auch lokale Vorhersagen interpretierbar gemacht werden. [11]

Wichtige Methoden, die diesen Ansatz nutzen sind LIME, Anchors, SmoothGrad, DeepRED und BETA. (Eine ausführlichere Beschreibung der einzelnen Methoden findet sich im Anhang 6.1)

LIME erzeugt ein Surrogat-Modell in der Bildklassifikation, indem es das Ausgangsbild in Teilbereiche unterteilt und die Relevanz der einzelnen Teilbereiche für die Vorhersage ermittelt. Zur Ermittlung der Relevanz werden Störungen (Ausblendungen) in das Bild eingebracht. Die Relevanzwerte ermöglichen die Ermittlung der Gewichte des Ersatzmodells. Die Methode ist sehr rechenaufwändig, kann aber für Laien verständliche Erklärungen erzeugen. [30, S. 12]

Anchors wurde zwei Jahre später vom selben Autor entwickelt. Die Methode entwickelt ein Surrogat-Modell über Entscheidungsregeln. Der Fokus liegt ähnlich wie bei

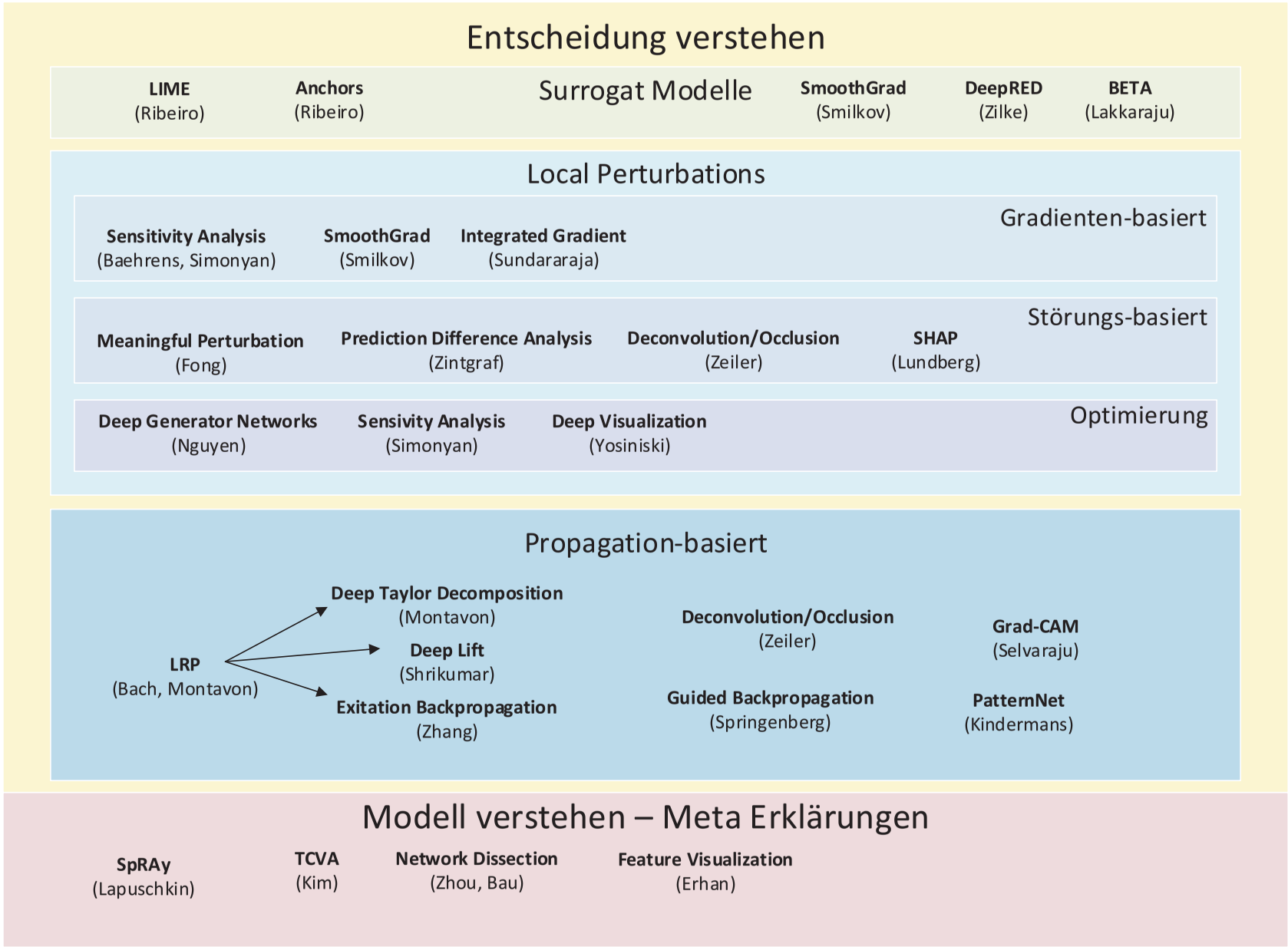


Abbildung 3.2: Zuordnung der Methoden zu den verschiedenen Ansätzen

LIME auf lokalen Entscheidungen und der unmittelbaren Umgebung davon. Auch hier werden Störungen verwendet, um Relevanzwerte zu ermitteln, die hilfreich für die Erzeugung der Entscheidungs-Regeln sind. Der Nachteil dieser Methode ist der hohe Rechenaufwand, der aber durch eine parallele Berechnung reduziert werden kann. [18, Kapitel 5.8]

SmoothGrad verwendet ähnlich wie LIME eine lokale Annäherung zur Erzeugung des Surrogat-Modells. Hier werden einzelne Datenpunkte inklusive Umgebung betrachtet, um näherungsweise einen Gradienten zu bestimmen. Die Methode kann daher auch den Gradienten-basierten Ansätzen zugeordnet werden. [30, S. 12f]

DeepRED verwendet ebenfalls einen regel-basierten Ansatz. In der Methode werden für jede Schicht eines Deep Neuronal Net (DNN) Zwischenregeln erzeugt. Die Regeln werden zu einem Regelsatz zusammengeführt und mit der Kombination dieser Regeln wird das Ursprungsmodell imitiert. [43, S. 1]

In der **BETA** Methode wird das zu erklärende Objekt in Teilbereiche aufgeteilt. Die Unterteilung sollte von einer Person mit Expertenwissen vorgenommen werden. Die Methode erzeugt Entscheidungsregeln, mit denen das Verhalten der Black-Box in den Teilbereichen simuliert wird. Diese Methode hat den Vorteil, dass das Verhalten des Black-Box Modells in ausgewählten Bereichen interaktiv untersucht werden kann. [16, S. 1])

3.2.2 Erklärung über lokale Störungen (local perturbation)

Ein anderer Ansatz zur Erzeugung von Erklärbarkeit, injiziert lokale Störungen in das komplexe KI-Modell und wertet die Reaktion aus. Bei diesem Vorgehen werden die Eingabewerte des Modells verändert, maskiert oder entfernt und mit diesen Änderungen wird ein erneuter Durchlauf durch das Neuronalen Netzes durchgeführt. Anschließend wird die Reaktion des Modells auf diese Änderungen analysiert. Dabei werden die Zuordnungen zwischen Eingabemerkmalen und Vorhersage neu berechnet. Durch die Zuordnung von Eingabemerkmalen und Ausgabenwerten kann die Interpretation des Modells erfolgen. Dieser Ansatz ist auf jedes Black-Box Modell anwendbar. [1, S.186]

Die Verwendung von lokalen Störungen zur Erzeugung von Erklärbarkeit hat jedoch Grenzen. Die Anzahl an Merkmalen, die gemeinsam gestört werden und die eingesetzte Störungsmethode, können die Ausgabe erheblich beeinflussen und die erzeugte Erklärung unbrauchbar machen. [1, S.186]

Die Störungs-basierte Erklärung hat zwei Vorteile: zum einen ist die Implementierung relativ einfach, zum anderen kann der Ansatz auf jede Modell-Architektur angewendet werden. Aber das Vorgehen hat seinen Preis: es ist sehr rechenaufwändig, da das Black-Box Modell mit einer großen Zahl an gestörten Eingabewerten durchlaufen werden muss, um eine Erklärung zu erzeugen.

Weiterhin besteht die Gefahr, dass derselbe Vorhersagewert von unterschiedlichen Eingangswerten erzeugt werden kann. Dieses Problem der Sättigung tritt auf, wenn ein Vorhersagewert bereits von einem Eingangsmerkmal erzeugt worden ist und ein anderes Eingangsmerkmal zu einem späteren Zeitpunkt denselben Vorhersagewert erzeugen möchte. Bei einem Erklärungsversuch durch ein Störungs-basiertes Verfahren kann durch das Sättigungsproblem die Relevanz einzelner Eingangsmerkmale für die Vorhersage falsch eingeschätzt werden. [13]

Es gibt verschiedene Methoden, um den Ansatz über lokale Störungen umzusetzen. Welche dieser Methoden für welchen Anwendungsfall besonders geeignet ist, wurde bisher kaum wissenschaftlich untersucht. Deshalb ist auch die Zuverlässigkeit der Vorhersagen dieser Methoden noch nicht abschließend geklärt. [1, S.186]

Der Ansatz über lokalen Störungen kann verschiedenartig umgesetzt werden. Es wird zwischen Gradienten-basierten Ansätzen und Störungs-basierten Ansätzen unterschieden. Hier werden verschiedene Methoden dieser beiden Ansätze vorgestellt. Gradienten-basierte Verfahren werden von [31] zu den lokalen Perturbations-Ansätzen gezählt, sie können jedoch auch als eigenständiger Ansatz betrachtet werden.[1, S.187]

3.2.2.1 Gradienten-basierte Ansätze

Gradient-basierte Ansätze verwenden die partielle Ableitung der Vorhersagefunktion um Erklärungen zu erzeugen. (Arras.2019, S.2) Dieses Vorgehen hat zwei Stärken. Zum einen ist der Rechenaufwand gering, da nur ein Vorwärts- und ein Rückwärtsdurchlauf des Netzwerkes benötigt wird, um eine Heatmap zu erstellen. Andererseits ist ihre Implementierung relativ einfach. Diese Verfahren haben jedoch verstärkt mit verrauschten Gradienten zu kämpfen. Dadurch ist es möglich, dass Methoden, die auf diesem Ansatz beruhen sehr empfindlich auf kleine Abweichungen bei den Eingabewerten reagieren. [1, S.183ff]

Aufgrund ihres relativ geringen Rechenaufwands, sind Gradienten-basierte Verfahren in der Regel schneller, als diejenige, die auf lokalen Störungen beruhen. [1, S.187]

Neben der SmoothGrad Methode, die sowohl mit einem Surrogat- als auch mit einem Gradienten-basierten Ansatz arbeitet und die bereits im Abschnitt der Surrogat-Modell vorgestellt wurde, werden hier die Methoden Sensivity Analysis und Integrated Gradients näher betrachtet.

In der **Sensivity Analysis** wird ermittelt, wie empfindlich die Ausgabe auf verschiedene Bereiche von Eingabewerten reagiert. Der Grad der Empfindlichkeit gibt Auskunft über die Relevanz der Eingabewerte für die Vorhersage. Die Empfindlichkeit wird über die Gradienten ermittelt und in einer Heatmap dargestellt. [20, S. 4]

Integrated Gradients erzeugt die Erklärung über verschiedene Kopien des Ausgangsbildes. Die Kopien werden mit unterschiedlichen Helligkeitswerten erstellt. Für diese Kopien werden jeweils die Gradienten berechnet. Aus dem Bezug des Mittelwerts aller Gradienten zum Originalbild wird die Relevanz der Eingabewerte für die Ausgabe ermittelt.[37] Der Rechenaufwand für diese Methode ist hoch, da der durchschnittliche Gradient numerisch ermittelt werden muss. [1, S. 182]

3.2.2.2 Störungs-basierte Ansätze (perturbation-based)

Störungs-basierte Ansätze versuchen die Vorhersage eines Modells hauptsächlich über die Reaktion des Modells auf eine injizierte lokale Störungen zu ermitteln. In dieser Arbeit werden die Methoden Meaningful Perturbation, Prediction Difference Analysis, Deconvolution und SHAP näher betrachtet.

In der **Meaningful Perturbation** Methode wird die Beziehung zwischen Eingabewerten und der Vorhersage des Modells durch interpretierbare Regeln abgebildet. Für die Erstellung relevanter Regeln werden minimale lokale Störungen in das System eingebracht und die Auswirkung davon auf die Ausgabe des Black-Box Modells untersucht. Das Verfahren ist sehr rechenaufwändig. [31, S. 146] und [6, S. 1f]

Prediction Difference Analysis kann zur Bilderkennung verwendet werden. Die Methode bringt Störungen in die Pixelumgebung eines analysierten Merkmals ein und analysiert die Auswirkungen auf das Vorhersageergebnis. Auf diese Weise werden die Relevanzwerte der einzelnen Merkmale ermittelt, die in einer Heatmap visualisiert werden können. Der Rechenaufwand der Methode ist hoch. [44, S. 1f]

Die **Deconvolution** Methode wird zur Bildklassifizierung in Convolutional Neural Networks (CNN) eingesetzt. Zur Erzeugung der Erklärung werden bedeutsame Bildbereiche systematisch sukzessiv abgedeckt. Durch die dadurch entstehende Änderung der Ausgabe werden die Relevanzwerte der Eingabemerkmale ermittelt. Über das Ausmaß der Änderungen können Rückschlüsse über die Relevanz der verdeckten Eingangswerte für die Vorhersage gezogen werden. [44, S. 2]

SHAP verwendet Shapley-Werte, die aus der kooperativen Spieltheorie bekannt sind, um die Vorhersage zu erklären. Über die Shapley-Werten werden die Beiträge, die die einzelnen Merkmale zur Vorhersage geleistet haben, ermittelt. Der Shapley-Wert ist der durchschnittlich erwartete Beitrag eines Merkmals zur Vorhersage, bezogen auf alle möglichen Kombinationen von Merkmalen (Koalitionen). Über die Shapley-Werte werden die Gewichte für die Vorhersage-Funktion ermittelt. Die Berechnung der Werte für alle möglichen Koalitionen ist extrem rechenaufwändig. Die Basis der Spieltheorie bietet eine solide theoretische Grundlage für diese Methode. [18, Kap. 5.10]

3.2.3 Optimierungs-basierte Ansätze

Die im vorherigen Abschnitt erläuterte Methode Meaningful Perturbation verwendet auch Optimierungsaspekte und könnte daher auch als Optimierungs-basiertes Verfahren betrachtet werden. Auch die Methoden Sensivity Analysis und LIME setzen Optimierung-Verfahren ein. Ebenfalls zur Kategorie der Optimierungs-basierten Ansätze werden die Methoden Deep Generator Networks und Deep Visualization gezählt, da diese Methoden Optimierungsverfahren einsetzen, um Erklärungen von maximalem Informationsgehalt zusammenzufassen. [30, S. 13]

Deep Generator Networks wird zur Klassifizierung von Bildern mit Neuronalen Netzen verwendet. Es wird zunächst ein Zufallsbild erzeugt und dann im Rückwärtsdurchlauf (backpropagation) durch das Neuronalen Netz ermittelt, wie jedes Pixel verändert werden muss, um die Aktivierung der Neuronen zu erhöhen, so dass sie dem Ausgangsbild entsprechen. Mithilfe von Aktivierungsmaximierung wird ein Prototyp erzeugt, der dem Vorhersageverhalten des zu erklärenden Black-Box Modelles möglichst nahekommt. [22, S. 1f]

Prinzipiell ist die Arbeitsweise von **Deep Visualization** und Deep Generator Networks ähnlich. Über Aktivierungsmaximierung werden die Eingabewerte ermittelt, die eine gute Annäherung an die Vorhersage des Modells erzeugen. Deep Visualization kombiniert dazu verschiedene Methoden miteinander, wie zum Beispiel Vorwärtsaktivierungswerte, Gradientenanstieg, dominierende Bilder und Rückwärtsdifferenzen. Die Methode ist robust gegenüber Änderungen von Parametern, kann aber rechenaufwändig sein. [40, S. 4]

3.2.4 Ausbreitungs-basierte Ansätze (propagation-based)

Ausbreitungs-basierte Ansätze werden oft zu Erzeugung von Erklärungen für tiefe Neuronale Netze eingesetzt. Wie schon bei den vorangegangenen Ansätzen ist die Beziehung zwischen Eingabemerkmalen und Vorhersage wichtig für das Verständnis der Entscheidungsfindung des untersuchten Modells.

Zur Ermittlung der Bedeutung der einzelnen Eingabemerkmale für die Vorhersage wird über einen Rückwärtsausbreitungsmechanismus (backpropagation) das Neuronale Netz rückwärts Layer für Layer von der erzeugten Vorhersage bis zu den Eingabemerkmalen durchlaufen. Dabei wird in jedem Layer die Bedeutung von einzelnen Neuronen für den Ausgabewert ermittelt. [13]

Ausbreitungs-basierte Ansätze agieren nicht völlig losgelöst vom zu erklärenden Modell, sondern sie nutzen dessen Struktur, um den Entscheidungsprozess transparent zu machen. [30, S. 13] Einige der Propagation-basierten Ansätze wie Layer-Wise-Relevance-Propagation (LRP), Guided Backpropagation und Excitation Backpropagation erzeugen die Erklärungen mit Hilfe von Informationen über die Backpropa-

gation Regeln und den Aktivierungen und Gewichten in den Convolutional Layern. [5, S. 159]

Die hier vorgestellten Methoden unterscheiden sich hauptsächlich dadurch, wie die Beiträge der einzelnen Neuronen für die vorherige Schicht ermittelt werden. Sie können über partielle Ableitungen wie bei Sensivity Analysis, Guided Backpropagtion und SmoothGrad erhoben werden. [13] Andere Methoden wie LRP und Deep Lift verwenden den Zerlegungsansatz, bei dem ein Zielneuronenaktivierungswert in seine Bestandteile zerlegt wird und eine Zuordnung der Bestandteile zur vorherigen Schicht stattfindet. [13]

Methoden, die nach dem Backpropagation Mechanismus arbeiten, benötigen für die Erstellung einer Erklärung für das Black-Box Modell nur einen oder wenige Durchläufe, daher ist der Rechenaufwand im Vergleich mit anderen Ansätzen gering.[13]

Ausbreitungs-basiert Methoden, die hier vorgestellt werden sind: Layerwise Relevance Propagation, Deep Taylor Decomposition, Deep Lift, Excitation Backpropagation, Guided Backpropagation, Grad-CAM und PatternNet.

Layerwise Relevance Propagation (LRP) ist eine Technik zur Ermittlung der Merkmale, die am stärksten zu der Vorhersage eines Neuronalen Netzes beigetragen haben. Nach einem erfolgten Vorwärtsdurchlauf durch das Neuronale Netz wird im Rückwärtsdurchlauf (backpropagation) das Relevanzmaß für jedes Merkmal ermittelt, indem auf jeden Layer des Netzes eine bestimmte Ausbreitungsregel angewendet wird. Die Relevanzwerte können in einer Heatmap dargestellt werden. So lässt sich ermitteln, welchen Beitrag jedes Merkmal zum Klassifizierungsergebnis geleistet hat. [2, S. 3] und [12, S. 138]

Deep Taylor Decomposition (DTD) ist verwandt mit LRP. Bereits in der wissenschaftlichen Arbeit zu LRP sind verschiedene Methoden für die Ermittlung der Ausbreitung der Relevanz in Neuronalen Netzen vorgesehen. DTD ist eine Variante, die dies ermöglicht. Bei dieser Methode wird die Funktion, die die Beziehung zwischen der Aktivierung von Layern und der Aktivierung von Knoten im Neuronalen Netz beschreibt, in partielle Ableitungen zerlegt. Durch diese partiellen Ableitungen kann die Relevanz-Ausbreitungsfunktion des Black-Box Modells approximiert werden. Zur Berechnung werden Taylor-Reihen verwendet.[1,2]

Deep Lift funktioniert ähnlich wie LRP. In der Backpropagation wird die Modell-Vorhersage auf die Eingabewerte zerlegt, um die Relevanz der Beiträge aller Neuronen des Netzwerkes zu ermitteln. Deep Lift setzt eine Referenzaktivierung zur Ermittlung der Referenzwerte der einzelnen Neuronen ein. Die Qualität der Methode hängt stark von den richtigen Referenzwerten ab. Um diese zu erzeugen, ist domänenspezifisches Wissen notwendig. [33, S. 1ff]

[1] **http://www.heatmapping.org/deeptaylor/**.
[2] **http://danshiebler.com/2017-04-16-deep-taylor-lrp/**.

Excitation Backpropagation geht von einem pyramidenförmigen Aufbau des neuronalen Netzes aus. Im ersten Durchlauf wird das Netz von unten nach oben durchlaufen und dabei die Stimuli des Netzes erkundet. Anschließend werden in einem umgekehrten Durchlauf die relevantesten Neuronen bestimmt. Die Beziehungen zwischen Vorhersagewerten und Eingabemerkmalen können in Heatmaps visuell dargestellt werden. [41, S. 1f]

Guided Backpropagation versucht Muster bei den Eingabewerten zu finden, die Einfluss auf die Ausgabe des Neuronalen Netzes gehabt haben. Dazu verändert die Methode die Architektur des Neuronalen Netzes: Die Max-Pooling Layer werden durch Convolutional Layer ersetzt. Alle negativen Gradienten werden auf null gesetzt. Dies ermöglicht die Erstellung von fokussierten Heatmaps. [36, S. 1f], [4, S. 5]

Grad-CAM ist nur für Convolutional Neural Nets (CNN) einsetzbar. Die Methode verwendet Class Activation Mapping (CAM), um die für eine Klassifizierung ausschlaggebenden Bereiche zu ermitteln. Über Gradienteninformationen aus der letzten Convolutional Schicht des CNN und erstellt die Methode eine grobe Feature Map, in der die für die Vorhersage wichtigen Bereiche hervorgehoben werden. Mit Hilfe der Methode können einzelne Objekte eines Bildes hervorgehoben werden. [32, S. 2f]

PatternNet wird eingesetzt, um die relevanten Bereiche eines Bildes zu ermittelt und um sie in einer Heatmap darzustellen. Die relevanten Bereiche werden hier als „Signal“ bezeichnet. Diese können zum Beispiele Objekte sein, die untersucht werden sollen. Die nicht-relevanten Bereiche und Rauschen wird entfernt. Mit Hilfe von **PatternAttribution** wird die Relevanz von den einzelnen Eingabemerkmalen (Pixeln) für die Klassifizierung herausgefunden. Dazu können verschieden Ansätze, wie zum Beispiel Backpropagation verwendet werden. [15, S. 1ff]

3.2.5 Erklärung des Modells

Während die bisher geschilderten Methoden ihren Fokus auf die Erklärbarkeit von Entscheidungen eines Modells gelegt haben, befassen sich die Methoden in diesem Abschnitt damit, das Modell zu verstehen. Diese Methoden versuchen allgemeine Muster aus dem Klassifizierungsverhalten des Modells zu extrahieren, indem einzelnen Entscheidungen zusammengefasst und anschließend analysiert werden.[30, S. 14] In diesem Abschnitt werden die Methoden Spectral Relevance Analysis (SpRAy), Feature Visualization, Network Dissection und Testing with Concept Activation Vectors (TCAV) vorgestellt.

Spectral Relevance Analysis (SpRAy) ist eine Weiterentwicklung von LRP. Mit dieser Methode werden zunächst Relevanzklassen für interessante Datensätze und Objektklassen über LRP bestimmt. Die Ergebnisse werden in Heatmaps dargestellt und anschließend über eine Spectralanalyse geclustert. Jedes Cluster entspricht einer

durch das Modell erlernten Vorhersagestrategie. Auf diese Weise werden die verwendeten Vorhersagestrategien für Objekte aus den erzeugten Clustern ermittelt. [17, S.3ff] Mit dieser Methoden lassen sich Schwachstellen in Datensätzen und Modellen erkennen. Es wird zum Beispiel erkannt, wenn eine Klassifizierung aufgrund der Metadaten eines Bildes vorgenommen wurde. [31, S. 14]

Feature Visualization arbeitet sich ebenso schrittweise durch das Neuronale Netz, um zu verstehen, wie das Modell sein Verständnis über das Ausgangsbild erstellt. Es wird erkundet, welche Eingabedaten die Ursachen für ein bestimmtes Verhalten (zum Beispiel eine Neuronen Aktivierung oder die endgültige Ausgabe) verantwortlich sind. Über eine Optimierungstechnik wird ein Verständnis darüber aufgebaut, wonach ein Modell sucht, dies können zum Beispiel Neuronen, Kanäle, Layer oder Klassenwahrscheinlichkeiten sein. [23]

Network Dissection ermöglicht zu verstehen, was in den einzelnen Layern eines vielschichtigen Convolutional Neural Networks (CNN) geschieht. Die Methode gleicht die Ausgabe jedes Layers mit visuellen semantischen Konzepten eines Vergleichsdatensatzes ab. Zum Vergleich wird der Broden-Datensatz eingesetzt, der viele Bilder und Konzepte enthält. Dieses Vorgehen ermöglicht die Zuordnung von Konzepten aus der realen Welt (zum Beispiel: unterschiedliche Materialien, Farben, Oberflächenstrukturen, Objekte, Szenen) zu jeder Schicht des CNN. Auf diese Weise werden die verborgenen Bereiche eines CNN interpretierbar gemacht werden und es können Erkenntnisse über den hierarchischen Aufbau des CNNs erlangt werden. [42, S. 1 und 12]

Testing with Concept Activation Vectors (TCAV) arbeitet nach demselben Prinzip wie Network Dissection. Die Ausgaben der verschiedenen Layer des Neuronalen Netzes werden mit Konzepten verglichen, die für Menschen verständlich sind, anschließend wird der Grad ihrer Übereinstimmung bewertet. Über die Bewertungen der ermittelten Konzepte wird die Wahrscheinlichkeit für die Klassifizierung des Ausgangsbildes berechnet. Die verschiedenen Konzepte werden im Neuronalen Netz als Vektoren dargestellt und die Wahrscheinlichkeit der Modellvorhersage wird über Richtungsableitungen bestimmt. [14, S. 1ff]

3.2.6 Methoden-Übersicht

Die folgenden Seiten bieten einen Überblick über die hier vorgestellten Methoden mit ihren jeweiligen Eigenschaften.

Methode	Autor	Community	Erklärungs-ansatz	Anwendungs-bereiche	ähnliche Methoden	Modell oder Entscheidung	lokal oder global	Modell-spezifisch, oder Modell-übergreifend	Performanz/ Effizienz	Bemerkungen
LIME	Ribeiro et al., 2016	University of Washington Microsoft Research	Surrogat, Perturbation, Optimierung	Bilder, Texte, Tabellen	SmoothGrad	Entscheidung	lokal	Modell-übergreifend	hoher Rechenaufwand	Erklärungen sind instabil
Anchors	Ribeiro et al., 2018	University of Washington Microsoft Research	Surrogat	Bilder, Texte, Tabellen	BETA	Entscheidung	lokal	Modell-übergreifend	hocheffizient da parallelisierbar	hohe Konfigurierbarkeit; in Trainingsphase rechenintensiv
SmoothGrad	Smilkov et al., 2017	Google Brain	Surrogat, Perturbation, Gradient	Bilderkennung	LIME, Sensivity Analysis	Entscheidung	lokal	Modell-spezifisch		durch Kombination mit anderen Sensivity Methoden schäfere Bilder
DeepRED	Zilke et al., 2016	TU Darmstadt	Surrogat	Buchstaben		Entscheidung	global	spezifisch für Neuronale Netze		noch in Entwicklungsphase, letzte Veröffentlichung: 2016
BETA	Lakkaraju et al., 2017	Stanford University Microsoft Research	Surrogat	Texterkennung	Anchors	Entscheidung	global	Modell-übergreifend		Nutzer können interaktiv interessante Bereiche des Black-Box Modells erkunden
Sensivity Analysis	Simonyan et al., 2014	University of Oxford	Perturbation, Gradient, Optimierung	Bilder, Texte	SmoothGrad, Integrated Gradients Deep Generator Networks	Entscheidung	lokal und global	Modell-übergreifend	Berechnung über Optimierung ist rechenaufwändig	erzeugt verauschte Visualisierungen
Integrated Gradients	Sundararajan et al., 2017	Google AI	Perturbation, Gradient	Bilder, Texte, chemische Modelle	DeepLift	Entscheidung	lokal	Modell-übergreifend	hoher Rechenaufwand	
Meaningful Perturbation	Fong & Vedaldi, 2017	University of Oxford	Perurbation, Optimierung	Bilderkennung	Deep Generator Networks	Entscheidung	lokal	Modell-übergreifend	hoher Rechenaufwand	

Abbildung 3.3: Übersicht über ausgewählte Methoden

Methode	Autor	Community	Erklärungs-ansatz	Anwendungs-bereiche	ähnliche Methoden	Modell oder Entscheidung	lokal oder global	Modell-spezifisch, oder Modell-übergreifend	Performanz/ Effizienz	Bemerkungen
Prediction Difference Analysis	Zintgraf et al., 2017	University of Amsterdam	Perturbation, Decomposition	Bilderkennung	Deconvolution Networks	Entscheidung	global	Modell-übergreifend	hoher Rechenaufwand, Berechung über GPUs möglich	
Deconvolution	Zeiler & Fergus, 2013	New York University	Perturbation, Deconvolution, Gradient, Propagation	Bilderkennung	Sensivity, Guided Backpropagation	Entscheidung	lokal	Modell-übergreifend		Kanten des Abdeckungs-rechtecks können Analyse-Ergebnis beeinflussen
SHAP	Lundberg & Lee, 2017	University of Washington	Perturbation	Bilder, Tabellendaten	LIME	Entscheidung	lokal und global	Modell-übergreifend	extrem hoher Rechenaufwand, langsam	höhere Effizienz durch Kombination mit anderen Methoden
Deep Generator Networks	Nguyen et al., 2016	Auburn University	Perturbation, Optimierung	Bilderkennung	Meaningful Perturbation Deep Visualization Sensivity	Entscheidung	lokal	Modell-übergreifend	Berechnung über Optimierung ist rechenaufwändig	
Deep Visualization	Yosinski et al., 2015	Cornell University Uber AI Labs	Perturbation, Gradient	Bilderkennung	Meaningful Perturbation Deep Generator Networks Sensivity	Entscheidung	lokal	spezifisch für Deep Neuronal Networks	hoher Rechenaufwand	robust gegenüber Änderungen
LRP	Bach et al, 2015	Fraunhofer TU Berlin	Propagation, Decomposition, Gradient	Bilder, Text, Sprache, Videos, Gesichts-erkennung, Spiele	Deep Taylor Decomposition, Excitation Backpropagation, Deep Lift	Entscheidung	lokal und global	spezifisch für Neuronale Netze	effiziente Berechnung	kann empfindlich für unwichtige Aspekte des Modells sein; teilweise numerisch instabil;
Deep Taylor Decomposition	Montavon et al., 2017	TU Berlin	Decomposition, Propagation	Bilderkennung	LRP	Entscheidung	lokal und global	spezifisch für Neuronale Netze		Methode ist stabil unter verschiedenen Architekturen
Deep Lift	Shrikumar et al., 2017	Stanford University	Gradient, Decomposition, Propagation	Bilder, Ziffern, DNA-Sequenzen	LRP, Integrated Gradients	Entscheidung	lokal	spezifisch für Neuronale Netze		kann empfindlich für unwichtige Aspekte des Modells sein

Abbildung 3.4: Übersicht über ausgewählte Methoden

Methode	Autor	Community	Erklärungs-ansatz	Anwendungs-bereiche	ähnliche Methoden	Modell oder Entscheidung	lokal oder global	Modell-spezifisch, oder Modell-übergreifend	Performanz/ Effizienz	Bemerkungen
Excitation Backpropagation	Zhang et al., 2016	Boston University Adobe Research	Gradient, Decomposition, Propagation	Bilder, Zuordnung von Texten zu Bildbereichen	LRP	Entscheidung	lokal	spezifisch für Convolutioal Neural Networks		Modell ist genau und generalisierbar
Guided Backpropagation	Springenberg et al., 2015	Uni Freiburg	Gradient, Propagation	Bilderkennung	Sensivity, Deconvolution	Entscheidung	lokal	spezifisch für Convolutioal Neural Networks		Methode erzeugt scharfe Bilder, ist anfällig für Störungen
Grad-CAM	Selvaraju et al., 2016	Virigina Tech	Gradient, Decomposition, Propagation	Bilderkennung		Entscheidung	lokal	spezifisch für Convolutioal Neural Networks		hohe Wiedergabetreue; Verzerrungen in Datensätzen werden erkannt
PatternNet, Pattern-Attribution	Kindermans et al., 2017	Google Brain TU Berlin	Optimization	Bilderkennung		Entscheidung	lokal	spezifisch für Neuronale Netze	effizient, aber in Trainingsphase höherer Rechenaufwand	
Spectral Relevance Analysis (SpRAy)	Lapuschkin et al., 2019	Fraunhofer TU Berlin	Meta-Erklärungen	BIlderkennung	basiert auf LRP	Modell	global	Modell-übergreifend		erkennt Schwachstellen in Modellen und Datensätzen
Feature Visualization	Erhan et al, 2009	Google AI	Meta-Erklärungen	Bilderekennung		Modell	global	spezifisch für Neuronale Netze		
Network Dissection	Zhou et al., 2017	Massachusetts Institute of Technology	Meta-Erklärungen	Bilderkennung		Modell	lokal	spezifisch für Concolutioal Neural Networks		Methode macht den Aufbau von CNNs transparent
Testing with Concept Activation Vectors (TCAV)	Kim et al, 2017	Google AI	Meta-Erklärungen, global Perturbation	Bilder, Audio, Video		Modell	global	spezifisch für Neuronale Netze		

Abbildung 3.5: Übersicht über ausgewählte Methoden

4 Diskussion der verschiedenen Erklärungsmodelle

Im vorherigen Kapitel wurden aktuell wichtige Ansätze vorgestellt, um KI-Modelle erklärbar zu machen. Zu jedem Ansatz wurden Methoden erläutert, die schwerpunktmäßig dem entsprechenden Ansatz folgen. Viele Methoden bestehen aus einer Kombination von verschiedenen Ansätzen. In diesem Kapitel werden die besprochenen Ansätze und Methoden miteinander verglichen und es wird auf ihre jeweiligen Vor- und Nachteile hingewiesen.

Surrogat Modelle

Surrogat Modelle kommen häufig zum Einsatz, wenn die Absicht besteht, tiefe Neuronal Netze durch Entscheidungsbäume oder Entscheidungsregeln zu ersetzen. Zum Beispiel kann die Methode DeepRED dazu eingesetzt werden, einen Entscheidungsbaum zu erstellen, um zu verstehen, wie eine Entscheidung zustande gekommen ist. Ein Entscheidungsbaum hat den Nachteil, dass er sehr groß werden kann, was die Nachvollziehbarkeit erschwert.

Um dieses Problem zu umgehen und die Komplexität der Erklärung zu begrenzen, verwendet die Surrogat-Methode Anchor Entscheidungsregeln für einzelne Instanz-Vorhersagen. In ähnlicher Weise setzt die Methode BETA eine begrenzte Anzahl von Entscheidungsregeln ein, um Teilbereiche eines komplexen Black-Box Modells zu imitieren. [13]

Perturbation-basierte Modelle

Wie im vorherigen Kapitel beschrieben, kann man Methoden, die mit gestörten Eingabewerten arbeiten, um Erklärbarkeit herstellen, je nach Schwerpunkt in Gradienten-basierte, Störungs-basierte, Propagations-basiert und Optimierungs-basierte Methoden differenzieren.

Die Gradienten-basierten Methoden betrachten das zu erklärende Neuronale Netz als Funktion und erzeugen die Erklärung über den Gradienten. [19, S. 253f] Sie errechnen

die Relevanz, die die jeweiligen Eingabemerkmale für die Vorhersage haben, in dem sie den Gradienten der Vorhersage in Bezug zu den Eingangswerte ermitteln. Dabei unterscheiden sich die einzelnen Methoden hauptsächlich dadurch, wie der Gradient durch die nicht-linearen Schichten des Neuronalen Netzes geführt wird. [13]

Gradienten-basierte Methoden haben den Vorteil, dass sie ebenso wie Propagations-basierte Methoden mit einem Vorwärts- und einem Rückwärtsdurchlauf durch das Neuronale Netz auskommen und dadurch relativ wenig rechnerischer Aufwand notwendig ist. [5, S. 158] Nachteilig wirken sich bei diesem Ansatz jedoch nicht-lineare Gradientenverläufe aus. In diesem Fall können missverständliche Relevanzwerte für die Merkmale das Ergebnis sein, wodurch es zu Verzerrungen kommen kann. Darüber hinaus ist es möglich, dass die Relevanz des Beitrags, den ein Merkmal zum Ergebnis geliefert hat, wegen bereits erfolgter Sättigung durch ein anderes Merkmal, unterschätzt wird. [33, S.1ff]

Dieser Effekt, die falsche Beurteilung der Relevanz einzelner Merkmale auf Grund von bestehender Sättigung, tritt auch bei Störungs-basierten Ansätzen auf. Viele Methoden, die Störung als Technik verwenden, setzen auf Optimierungsverfahren. [5, S. 158] Die Methoden sind rechnerisch nicht effizient und langsam, da für jede ins Neuronale Netz eingebrachte Störung ein neuer Vorwärtsdurchlauf durch das Netz notwendig ist. [33, S.1ff]

Die Ergebnisse von Störungs-basierten Verfahren sind nicht immer stabil [33, S.1ff] Um stabilere Heatmaps zu ermöglichen, setzt die Integrated Gradients-Methode Gradienten-Berechnungen ein. [13] Der Vorteil von störungs-basierten Ansätzen ist ihre leichte Implementierbarkeit und ihre Flexibilität hinsichtlich der zu erklärenden Modellarchitektur. [13]

Propagations-basierte Modelle

Im Gegensatz zu Störungs-basierten Vorgehensweisen sind Methoden die über Propagation arbeiten recheneffizienter, da sie nur wenige Durchläufe durch das Netz benötigen, um eine Erklärung zu erzeugen. Dieser Ansatz funktioniert jedoch nur, wenn das zu erklärende Modell ein Neuronales Netz ist.

Einige Propagation-basierte Verfahren verwenden Gradienten zur Erzeugung der Erklärung. Es gibt verschiedene Methoden, die diese beiden Ansätze kombinieren, zum Beispiel Sensivity Analysis und Deconvolution. Die Methoden, die beide Ansätze kombinieren unterscheiden sich hauptsächlich dadurch, wie sie den Gradienten in der Backpropagation durch die nicht-linearen Schichten des Neuronalen Netzes führen. [33, S.2]

In der Sensivity Analysis wird der Gradient im Rückwärtsdurchlauf beim Eintritt in die nicht-lineare Schicht auf null gesetzt, wenn er im Vorwärtsdurchlauf beim Eintritt in diese Schicht bereits negativ war. Die Deconvolutional-Methode verwendet die

Informationen aus dem Vorwärtsdurchlauf nicht, sondern setzt den Gradienten auf null, wenn er beim Rückwärtsdurchlauf beim Eintritt in die nicht-lineare Schicht negativ ist. [33, S.2]

Die Methode Guided Backpropagation kombiniert das Vorgehen von Sensivity Analysis und Deconvolutional. Alle drei Methoden sind anfällig für das Sättigungsproblem und haben für das Durchleiten des Gradienten durch die nicht-linearen Schichten keine befriedigende Lösung. Die Grad-CAM-Methode umgeht dieses Problem, indem sie die Relevanz der Merkmale über den letzten Convolutional Layer ermittelt. [33, S.2]

Viele propagations-basierte Methoden haben [5, S. 159] den Nachteil, das eine Änderung der Backpropagation-Regeln erforderlich ist (Guided Backpropagation, Deconvolution, Network Dissection), das bedeutet, das ein Eingriff in das zu untersuchende Modell notwendig ist. Andere Methode, wie zum Beispiel LRP, Grad-CAM, Excitation Backpropagation und Network Dissection benötigen einen Zugriff auf die Zwischenschichten des Modells.

Layerwise Relevance Propagation (LRP) ist eine Methode, die in der Fachwelt viel Aufmerksamkeit hervorgerufen hat. Es existieren mit DeepLift, Excitation Backpropagation, Gradient Times Input und Deep Taylor Decomposition einige Methoden, die sehr ähnlich sind. Gegenüber Gradienten-basierten Ansätzen, hat LRP den Vorteil, dass erzeugte Heatmaps leichter zu interpretieren sind. Wenn Gradienten-basierte Ansätzen verwendet werden, können die Ergebnisse stark verrauscht und dadurch schwer zu interpretieren sein. Ein weiterer Vorteil von LRP gegenüber Gradienten-basierten Ansätzen, ist, dass LRP auch die Merkmale in der Heatmap kennzeichnet, die einen negativen Einfluss auf die Vorhersage haben. Gradienten-basierte Ansätze können dies nicht. [20, S. 6]

Nach Einschätzung von [4, S. 2] weisen Gradienten-basierte Verfahren nur darauf hin, wie empfindlich das Modell auf Änderungen reagiert, während LRP die eigentliche Ursache für eine erfolgte Klassifizierung aufzeigt. Nach Einschätzung des Autors ist LRP den Gradienten-basierten Ansätzen und den Methoden, die auf Entfaltung beruhen, wie zum Beispiel Deconvolution und Guided Backpropagation, überlegen.

Der Autor von [13] sieht jedoch auch Kritikpunkte bei den Propagtions-basierten Ansätzen. Nach seiner Einschätzung sind die Erklärungen von LRP und DeepLift nicht immer zuverlässig. Die von DeepLift erzeugten Erklärungen sind stark von dem vom Nutzer gewählten Referenzpunkt abhängig und die Ergebnisse von LRP können numerisch instabil sein.

5 Schlussbetrachtung

5.1 Eigene Bewertung des Masterprojekts

In dieser Arbeit wurden einige Ansätzen und Methoden vorgestellt, die mehr Transparenz in das Zustandekommen von Vorhersagen und die Arbeitsweise von verschiedenen Methoden der KI bringen können. Bei der Beschreibung der einzelnen Methoden fällt auf, dass derzeit noch kein allgemein akzeptierter Kriterienkatalog existiert, mit denen diese Methoden beschrieben werden können. Es gibt zwar verschiedene Ansätze von Autoren, die wichtige Kriterien beschreiben [27], [18], einige wurden auch in dieser Arbeit in Abschnitt 2.4 erläutert. Jedoch gibt es bis heute keinen allgemein anerkannten Standard, der zur Beschreibung der Methoden angewendet werden kann.

Da bisher wenige wissenschaftliche Arbeiten sich einem vergleichenden Überblick der wichtigsten Methoden und Ansätze gewidmet haben, ist es auf Basis von reiner Literaturrecherche schwer möglich, die einzelnen Methoden mit einem einheitlichen Kriterienkatalog zu beschreiben. Die dazu notwendigen Informationen finden sich weder in den wissenschaftlichen Arbeiten, die die Methoden beschreiben, noch in den wenigen Arbeiten, die einen allgemeinen Überblick geben.

Um die einzelnen Methoden mit einem einheitlichen Kriterienkatalog zu bewerten, müssten alle relevanten Methoden im Praxiseinsatz in Bezug auf die Einhaltung von vorher definierten Kriterien analysiert werden. Dies war jedoch nicht Aufgabe dieses Masterprojekts und wäre auch in dem zeitlichen Rahmen des Projekts nicht möglich gewesen.

Ziel des Projekts war, einen Überblick über den aktuellen Stand von relevanten Methoden zu erarbeiten, die Modelle der KI erklärbar machen können. Die anschließende Masterthesis soll auf diesen, im Projekt gesammelten Informationen und gewonnen Einsichten, aufbauen. Hier wird dann die Zeit sein, einzelne Methoden genauer zu untersuchen und den Grad der Eignung für konkrete Aufgaben zu prüfen.

5.2 Zukünftige Entwicklung

Obwohl in den letzten Jahren erhebliche Fortschritte hinsichtlich der Erklärbarkeit von KI gemacht wurden, stehen die Wissenschaftler weiter vor großen Herausforderungen, sowohl in der theoretischen Grundlage, als auch in der methodischen Umsetzung. Derzeit existiert noch keine allgemein anerkannte Theorie über erklärbare Künstliche Intelligenz. Es gibt keine Definitionen, die von den Forschern auf diesem Gebiet als allgemeingültig gesehen werden und kein Rahmenwerk, für die Beschreibung von neuen Methoden. Somit fehlt auch ein Maß zur Bewertung der

Qualität von einzelnen Methoden. Eine Vergleichbarkeit von verschiedenen Methoden ist daher schwer möglich. [31, S. 16f und S.240]

Die Erklärungen, die die hier beschriebenen Methoden erzeugen, sind meist von visueller Natur und bestehen oft darin, dass die für eine Vorhersage relevanten Merkmale zum Beispiel in Heatmaps hervorgehoben werden. Eine Beziehung zwischen relevanten Pixeln und dem eigentlichen Objekt kann durch die Erklärungen noch nicht hergestellt werden. Sie sind daher noch von einem geringen Abstraktionsgrad und können keine abstrakten Konzepte erklären, wie zum Beispiel ganze Objekte, Ansammlungen von Objekten oder gar die Interaktion zwischen verschiedenen Objekten.[31, S. 16f]

Erklärungsmodelle können zwar die Relevanz von einzelnen Merkmalen für eine Vorhersage ermitteln, es bleibt jedoch oft unklar, ob ein einzelnes Merkmal alleine für einen Ausgabewert eines Modells verantwortlich ist oder ob die Kombination von verschiedenen Merkmalen ein Ergebnis erzeugt hat. In Zukunft wird es wahrscheinlich Methoden geben, die abstraktere Erklärungen erzeugen können, die nicht nur für Experten mit Machine Learning-Kenntnissen verständlich sind.

Um ein besseres Verständnis über den Vorgang der Entscheidungsfindung von Modellen der KI zu erlangen, wäre es sinnvoll, dass Forscher aus verschiedenen Bereichen interdisziplinär zusammenarbeiten. So könnten Experten aus den Bereichen Machine Learning, Mensch-Computer-Interaktion und Psychologen zusammen mit Menschen, die sich in der Domäne auskennen, in der das KI-Modell eingesetzt werden soll, ein Modell entwickeln, dass allgemein-verständliche Erklärungen erzeugt.

Ein anderer Ansatz für mehr Transparenz könnte sein, einen stärkeren Fokus in der Forschung auf die Entwicklung von Modellen zu legen, die von Natur aus erklärbar sind, statt Methoden zu entwickeln, die Black-Box Modelle erklären können. Die Autorin [28, S. 1ff] vertritt diese Einstellung. Ihrer Meinung nach hat sich die Wissenschaft in den letzten Jahren zu sehr darauf fokussiert, Erklärungen für Black-Box Modelle zu erschaffen und sich zu wenig damit befasst, Modelle zu entwickeln, die in sich erklärbar sind. Viele Aufgaben, die von Black-Box Modellen erledigt werden, könnten auch transparente Modelle erledigen. Ein zweites Modell und damit zusätzliche Fehlerquellen würden wegfallen. Daher plädiert sie dafür, Black-Box Modelle nur dann einzusetzen, wenn für dieselbe Aufgabe kein transparentes Modell vorhanden ist. Sie erwartet von komplexeren Modellen nicht notwendigerweise genauere Ergebnisse, im Vergleich mit einfachen Modellen.

6 Literaturverzeichnis

[1] Ancona, Marco u. a. „Gradient-Based Attribution Methods“. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Hrsg. von Samek, Wojciech u. a. Cham: Springer International Publishing, 2019, S. 169–191. ISBN: 978-3-030-28953-9.

[2] Arras, Leila u. a. *Evaluating Recurrent Neural Network Explanations*. 2019. URL: https://arxiv.org/abs/1904.11829v3 (besucht am 31. 10. 2019).

[3] Arras, Leila u. a. „Explaining and Interpreting LSTMs“. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Hrsg. von Samek, Wojciech u. a. Cham: Springer International Publishing, 2019, S. 211–242. ISBN: 978-3-030-28953-9.

[4] Böhle, Moritz u. a. „Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification“. In: *Frontiers in aging neuroscience* 11 (2019), S. 194. ISSN: 1663-4365. DOI: 10.3389/fnagi.2019.00194. URL: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194/full (besucht am 31. 10. 2019).

[5] Fong, Ruth und Vedaldi, Andrea. „Explanations for Attributing Deep Neural Network Predictions“. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Hrsg. von Samek, Wojciech u. a. Cham: Springer International Publishing, 2019, S. 149–167. ISBN: 978-3-030-28953-9.

[6] Fong, Ruth und Vedaldi, Andrea. „Interpretable Explanations of Black Boxes by Meaningful Perturbation“. In: (2017), S. 3449–3457. URL: https://arxiv.org/abs/1704.03296v3 (besucht am 29. 10. 2019).

[7] Ghosh, Rohit und Shubham Jain. *Visualizing Deep Learning Networks*. 2017. URL: http://blog.qure.ai/notes/visualizing_deep_learning.

[8] Gill, Navdeep und Hall, Patrick. *Introduction to Machine Learning Interpretability*. Sebastopol, CA: O'Reilly Media, Inc, 2018. ISBN: 9781492033158.

[9] Glander, Shirin. *Künstliche Intelligenz und Erklärbarkeit _ Informatik Aktuell*. 2018. URL: https://www.informatik-aktuell.de/betrieb/kuenstliche-intelligenz/kuenstliche-intelligenz-und-erklaerbarkeit.html (besucht am 09. 11. 2019).

[10] Guidotti, Riccardo u. a. *A Survey Of Methods For Explaining Black Box Models*. 2018. URL: https://arxiv.org/abs/1802.01933v3.

[11] Hall, Patrick; Phan, Wen und Ambati, SriSatish. *Ideas on interpreting machine learning: Mix-and-match approaches for visualizing data and interpreting machine learning models and results.* 2017. URL: https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning (besucht am 14. 10. 2019).

[12] Holzinger, Andreas. „Interpretierbare KI: Neue Methoden zeigen Entscheidungswege künstlicher Intelligenz auf". In: *ct Magazin* 22 (2018), S. 136–141.

[13] Khaleghi, Bahador. *The How of Explainable AI: Post-modelling Explainability.* 2019. URL: https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f (besucht am 31. 10. 2019).

[14] Kim, Been u. a. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).* 2017. URL: https://arxiv.org/abs/1711.11279v5 (besucht am 06. 11. 2019).

[15] Kindermans, Pieter-Jan u. a. *Learning how to explain neural networks: PatternNet and PatternAttribution.* 2017. URL: https://arxiv.org/abs/1705.05598v2 (besucht am 05. 11. 2019).

[16] Lakkaraju, Himabindu u. a. *Interpretable & Explorable Approximations of Black Box Models.* 2017. URL: https://arxiv.org/abs/1707.01154v1 (besucht am 24. 10. 2019).

[17] Lapuschkin, Sebastian u. a. „Unmasking Clever Hans predictors and assessing what machines really learn". In: *Nature communications* 10.1 (2019), S. 1096. URL: https://www.ncbi.nlm.nih.gov/pubmed/30858366 (besucht am 05. 11. 2019).

[18] Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Leanpub, 2019. URL: https://christophm.github.io/interpretable-ml-book/ (besucht am 06. 11. 2019).

[19] Montavon, Grégoire. „Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Hrsg. von Samek, Wojciech u. a. Cham: Springer International Publishing, 2019, S. 253–265. ISBN: 978-3-030-28953-9.

[20] Montavon, Grégoire; Samek, Wojciech und Müller, Klaus-Robert. „Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73 (2018), S. 1–15. ISSN: 10512004. DOI: 10.1016/j.dsp.2017.10.011.

[21] Montavon, Grégoire u. a. „Explaining NonLinear Classification Decisions with Deep Taylor Decomposition". In: *Pattern Recognition* 65 (2017), S. 211–222. ISSN: 00313203. URL: https://www.sciencedirect.com/science/article/pii/S0031320316303582?via%3Dihub (besucht am 31. 10. 2019).

[22] Nguyen, Anh u. a. *Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.* 2016. URL: https://arxiv.org/abs/1605.09304v5 (besucht am 30. 10. 2019).

[23] Olah, Chris; Mordvintsev, Alexander und Schubert, Ludwig. *Feature Visualization: How neural networks build up their understanding of images*. 2017. URL: https://distill.pub/2017/feature-visualization/#enemy-of-feature-vis (besucht am 05. 11. 2019).

[24] Ribeiro, Marco Tulio. *Local Interpretable Model-Agnostic Explanations (LIME): An Introduction*. 2016. URL: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime (besucht am 07. 11. 2019).

[25] Ribeiro, Marco Tulio; Singh, Sameer und Carlos Guestrin. *Anchors: High Precision Model-Agnostic Explanations*. Hrsg. von University of Washington. 2018. URL: https://homes.cs.washington.edu/~marcotcr/aaai18.pdf.

[26] Ribeiro, Marco Tulio; Singh, Sameer und Guestrin, Carlos. „"Why Should I Trust You?"“ In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. Hrsg. von Krishnapuram, Balaji u. a. New York, New York, USA: ACM Press, 2016, S. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778.

[27] Robnik-Šikonja, Marko und Bohanec, Marko. „Perturbation-Based Explanations of Prediction Models“. In: *Human and Machine Learning*. Hrsg. von Zhou, Jianlong und Chen, Fang. Cham: Springer International Publishing, 2018, S. 159–175. ISBN: 978-3-319-90402-3.

[28] Rudin, Cynthia. „Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead“. In: *Nature Machine Intelligence* (2019). URL: http://arxiv.org/pdf/1811.10154v3 (besucht am 26. 11. 2019).

[29] Samek, Wojciech. *Interpreting and Explaining Deep Models in Computer Vision: IAPR Summer School on Machine and Visual Intelligence*. Vico Equense, Naples, Italy, 2018. URL: http://iphome.hhi.de/samek/pdf/VISMACSummerSchool2018.pdf.

[30] Samek, Wojciech und Müller, Klaus-Robert. „Towards Explainable Artificial Intelligence“. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Hrsg. von Samek, Wojciech u. a. Cham: Springer International Publishing, 2019, S. 5–22. ISBN: 978-3-030-28953-9.

[31] Samek, Wojciech u. a., Hrsg. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Bd. 11700. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-28953-9. DOI: 10.1007/978-3-030-28954-6.

[32] Selvaraju, Ramprasaath R. u. a. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. 2016. URL: https://arxiv.org/abs/1610.02391v3 (besucht am 04. 11. 2019).

[33] Shrikumar, Avanti; Greenside, Peyton und Kundaje, Anshul. *Learning Important Features Through Propagating Activation Differences*. 2017. URL: https://arxiv.org/abs/1704.02685v1 (besucht am 31. 10. 2019).

[34] Smilkov, Daniel u. a. *SmoothGrad*. URL: https://pair-code.github.io/saliency/ (besucht am 22. 10. 2019).

[35] Smilkov, Daniel u. a. *SmoothGrad: removing noise by adding noise*. 2017. URL: https://arxiv.org/abs/1706.03825v1.

[36] Springenberg, Jost Tobias u. a. *Striving for Simplicity: The All Convolutional Net*. 2015. URL: https://arxiv.org/abs/1412.6806v3 (besucht am 04. 11. 2019).

[37] Sundararajan, Mukund; Taly, Ankur und Qiqi Yan. *Attributing a deep network's prediction to its input features: The Unofficial Google Data Science Blog*. 2017. URL: http://www.unofficialgoogledatascience.com/2017/03/attributing-deep-networks-prediction-to.html (besucht am 28. 10. 2019).

[38] Sundararajan, Mukund; Taly, Ankur und Yan, Qiqi. *Axiomatic Attribution for Deep Networks*. 2017. URL: http://arxiv.org/pdf/1703.01365v2.

[39] Webb, Amy und Pyka, Petra. *Die großen Neun: Wie wir die Tech-Titanen bändigen und eine künstliche Intelligenz zum Wohle aller entwickeln können : Microsoft, Alibaba, IBM, Google, Amazon, Facebook, Tencent, Apple, Baidu*. 2019. ISBN: 9783864706387.

[40] Yosinski, Jason u. a. *Understanding Neural Networks Through Deep Visualization*. 2015. URL: https://arxiv.org/abs/1506.06579v1 (besucht am 30. 10. 2019).

[41] Zhang, Jianming u. a. *Top-down Neural Attention by Excitation Backprop*. 2016. URL: https://arxiv.org/abs/1608.00507v1 (besucht am 04. 11. 2019).

[42] Zhou, Bolei u. a. *Interpreting Deep Visual Representations via Network Dissection*. 2017. URL: https://arxiv.org/abs/1711.05611v2 (besucht am 06. 11. 2019).

[43] Zilke, Jan Ruben; Mencia, Eneldo Loza und Janssen, Frederik. „DeepRED – Rule Extraction from Deep Neural Networks“. In: *Discovery Science*. Cham: Springer International Publishing, 2016, S. 457–473.

[44] Zintgraf, Luisa M. u. a. *Visualizing Deep Neural Network Decisions: Prediction Difference Analysis*. 2017. URL: https://arxiv.org/abs/1702.04595v1 (besucht am 29. 10. 2019).

6.1 Surrogat-Modelle

Local Interpretable Model-Agnostic Explanations (LIME)

Referenz: *Ribeiro et al.: „Why Should I Trust You?“ Explaining the Predictions of Any Classifier, 2016*

Modellbeschreibung: LIME erzeugt Erklärbarkeit für ein KI-Modell über ein leichter interpretierbares Ersatzmodell. Dieses kann zum Beispiel ein lineares Modell sein. Zur Erstellung des Ersatzmodells fokussiert sich die Methode auf die unmittelbare Umgebung des Bereichs, der für die Vorhersage ausschlaggebend ist und versucht hier eine lokale Annäherung. Eine lokale Approximation ist leichter zu ermitteln, als eine globale.

Die lokale Annäherung in der Bildklassifikation wird erreicht, indem das Surrogat-Modell über das Verhalten des Black-Box Modells nach dem Einbringen von Störungen (z.B. durch das Ausblenden von Bildbereichen) trainiert wurde. Dabei werden die gestörten Bilder hinsichtlich des Grads der Ähnlichkeit zu der zu erklärenden Instanz gewichtet.

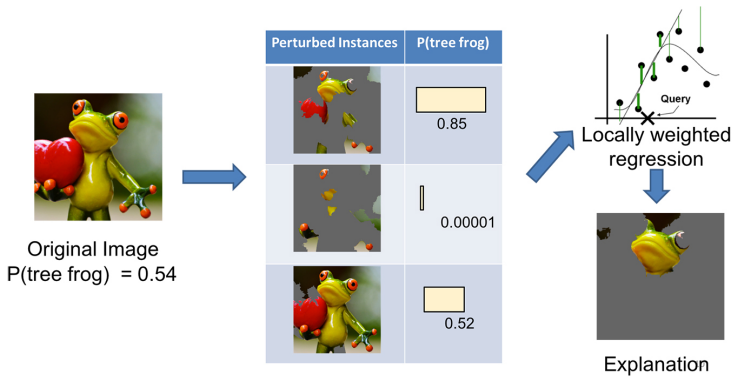


Abbildung 6.1: Bildklassifizierung mit LIME [24]

Das zu klassifizierende Bild wird in kleinere zusammenhängende Bildbereiche (Superpixel) unterteilt und während des Klassifizierungsprozesses werden jeweils verschiedene Bildbereiche mit einem grauen Rechteck verdeckt. So kann ermittelt werden, welche Bildbereiche für die Vorhersage wichtig sind und welche weniger wichtig.

Anwendungsbereiche: Bild- und Textklassifikation, Tabellen; [18, Kap. 5.7]

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:

- die durch LIME erzeugten Erklärungen sind kurz und kontrastreich und für Laien verständlich;
- vielfältige Anwendungsmöglichkeiten: Bilder, Texte und Tabellendaten;[18, Kap. 5.7]

Nachteile:
- die Erklärungen sind nicht immer stabil;
- die durch LIME erzeugten Erklärungen können für Experten zu oberflächlich sein;
- bei Tabellendaten ist die Definition der unmittelbaren Umgebung eines Datenpunktes schwierig; [18, Kap. 5.7]
- hoher Rechenaufwand; [31, S. 12]

Implementierung:
- lime[a] und skater[b] (Python);
- lime package[c] und iml package[d] (R);

[a]https://github.com/marcotcr/lime
[b]https://github.com/oracle/Skater
[c]https://cran.r-project.org/web/packages/lime/index.html
[d]https://cran.r-project.org/web/packages/iml/index.html

Anchors

Referenz: *Ribeiro et al.: Anchors: High Precision Model-Agnostic Explanations, 2018*

Modellbeschreibung: Anchors erklärt die Vorhersagen eines Black-Box Modells über Entscheidungsregeln, mit denen die Vorhersage verankert wird. Die Entscheidungsregeln werden als Anker bezeichnet und bestehen aus „wenn-dann-Regeln“. Aufgabe der Regeln ist die hinreichende lokale Verankerung der Vorhersage, so dass eine Änderung von Merkmalswerten die Vorhersage nicht mehr beeinflussen können. [25, S. 1] Zur Ermittlung der Anker wird für jede zur erklärende Instanz Nachbarwerte betrachtet oder Störungen erzeugt und ausgewertet.[18, Kapitel 5.8]

Globale regelbasierte Erklärungsansätze sind meistens sehr rechenaufwändig und können die Komplexität von tiefen neuronalen Netzen oft nicht gut abbilden. Anchors versucht diesen Nachteil durch die Verwendung von lokalen Regeln für Instanz-Vorhersagen zu vermeiden.[13]

Anwendungsbereiche: Anchors kann zur Lösung von Klassifizierungsproblemen in Data Mining und zur Mustererkennung eingesetzt werden.[8, S. 29] Das Verfahren ist zur Klassifikation von Texten, Tabellen und Bildern geeignet.[25, S. 2f]

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Model-übergreifend;

Vorteile:
- die Ausgabe ist auch für Laien gut zu verstehen;
- Anchors funktioniert auch, wenn die Modellvorhersagen in der Nachbarschaft der Instanz nicht-linear oder komplex sind;
- das Verfahren ist hocheffizient, da es parallelisierbar ist; [18, Kapitel 9.8]

Nachteile:
- hohe Konfigurierbarkeit (d.h. die vielen Einstellungsmöglichkeiten erschweren das Erlangen von aussagekräftigen Ergebnissen);
- zur Erzeugung von Ankern sind viele Aufrufe des Black-Box Modells notwendig; [18, Kapitel 9.8]

Implementierung:
- anchor (Python);
- anchorsOnR (R);
- AnchorJ (Java);

SmoothGrad

Referenz: *Smilkov et al.: SmoothGrad: removing noise by adding noise, 2017*

Modellbeschreibung: SmoothGrad funktioniert ähnlich wie LIME. Es wird ein Modell zur lokalen Annäherung der Black-Box Funktion erstellt. Die Methode verwendet Datenpunkte aus der Umgebung der Eingabewerte und berechnet auf dieser Basis eine Annäherung an den Gradienten für diesen Bereich. Außer den Zugriff zu den Gradienten benötigt SmoothGrad keine weiteren internen Informationen über das Black-Box-Modell. Die Methode ist eine Kombination zwischen Surrogat-Modell und Gradienten-basierten Ansatz. [30, S. 12f]

Die SmoothGrad-Technik kann auch zur visuellen Verbesserung von Gradienten-basierten Sensivitäts-Masken eingesetzt werden. Nach der Anwendung der Methode wird erkennbar, wie sich kleine Änderungen an Pixeln im Ausgangsbild auf die Vorhersage des Modells auswirken. Das Vorgehen ist wie folgt: Erstellung von vielen Kopien des zu analysierenden Bildes, die jeweils unterschiedlich verrauscht werden. Anschließend wird die Vorhersagefunktion auf alle Kopien angewendet. Dann wird der Mittelwert über die Ergebnisverläufe gebildet. [34, 35, S. 1]

Anwendungsbereich: Bilder;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Model-spezifisch;

Vorteile:

- SmoothGrad kann zu einer höheren Schärfe in Sensivity Masks beitragen [35, S. 5]

Implementierung: saliency (Python)[a]

[a]https://github.com/pair-code/saliency

;

DeepRED (Rule Extraction from Deep Neural Networks)

Referenz: *Zilke et al.: DeepRED – Rule Extraction from Deep Neural Networks, 2016*

Modellbeschreibung: Mit DeepRED können Regeln aus tiefen Neuronalen Netzen extrahiert werden. Da das regelbasierte Lösen von Klassifizierungsproblemen dem menschlichen Vorgehen ähnelt, ist diese Vorgehensweise gut geeignet um Entscheidungsprozesse für Menschen verständlich zu machen. Bisher wurden regelbasierte Erklärungsansätze nur für kleine Neuronale Netze mit einem Layer angewendet, DeepRED möchte den Ansatz auch zur Erklärung von einem komplexen Deep Neuronal Net (DNN) mit mehreren Layern zu nutzen.

DeepRED erzeugt Zwischenregeln für jede Schicht eines DNN. Die Zwischenregeln werden anschließend zu einem Regelsatz zusammengeführt, der den Entscheidungsprozess des DNNs imitiert. [43, S. 1]

Anwendungsbereiche: Erkennung von Buchstaben; weitere Anwendungsbereiche sind nicht bekannt;

Reichweite der Erklärungen: global;

Modell-spezifisch oder Modell-übergreifend: spezifisch für Neuronale Netze;

Vorteile:
- regelbasierte Erklärungen sind für Menschen verständlich;

Nachteile:
- das Methode befindet sich anscheinend noch in der Entwicklungsphase (nach der Veröffentlichung im Jahr 2016 wurden keine weiteren Informationen zu der Methode mehr publiziert);

Implementierung: keine veröffentlichte Implementierung bekannt;

BETA
(Black Box Explanations through Transparent Approximations)

Referenz: *Lakkaraju et al.: Interpretable & Explorable Approximations of Black Box Models, 2017*

Modellbeschreibung: BETA hilft dem Anwender die Komplexität von Klassifikationsentscheidungen zu mindern, indem Teilbereiche (subspaces) im Merkmalsraum definiert werden. Zur Definition der Teilbereiche sind Sachkenntnisse eines Experten notwendig, d.h. der Experte muss sich mit dem Kontext der Entscheidung auskennen. Mit dem Modell kann ein Experte interaktiv das Verhalten des Black-Box Modells in den für ihn interessanten Teilbereichen verstehen. [12, S. 137f]

BETA verwendet Entscheidungsregeln (wenn-dann-Regeln), um das Verhalten des Black-Box Modells in den definierten Teilbereichen zu erfassen. Durch transparente Entscheidungsprozesse verhindert BETA, das die Entscheidungsregeln nicht überlappende Bereiche des Merkmalsraums erklären. [16, S. 1f]

Anwendungsbereiche: Texte;

Reichweite der Erklärungen: global;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile: (da keine anderen Informationen vorlagen, wurden die Vorteile aus der wissenschaftlichen Arbeit zu der Methode übernommen [16, S. 1])
- das Black-Box Modell wird durch BETA hinsichtlich seiner Eindeutigkeit, Wiedergabetreue und Interpretierbarkeit optimiert;
- der Nutzer kann das Modellverhalten hinsichtlich seiner Präferenzen untersuchten;
- BETA erzeugt leicht verständliche und exakte Surrogate;

Nachteile:
- zur Festlegung der Teilbereiche ist Expertenwissen notwendig; [12, S. 137f]

Implementierung: keine publizierte Implementierung bekannt;

6.2 Erklärungen über lokale Störungen

6.2.1 Gradienten-basierte Ansätze

Sensivity Analysis

Referenz: *Simonyan et al.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2014*

Modellbeschreibung: Diese Methode setzt den Gradienten der Ausgabe des Black-Box Modells in Beziehung zu den Eingabewerten. Eingabewerte sind die Pixel in der Klassifikation von Bildern. Die Ergebnisse von der Sensivity Analysis-Methode können in einer Heatmap visuell dargestellt werden.

Ziel der Methode ist, zu ermitteln, wie empfindlich die aktuelle Ausgabe in Bezug auf verschiedene Teile der Eingabe reagiert. Je größer die Empfindlichkeit ist, desto mehr Relevanz haben die Eingabewerte bei der Erzeugung der Vorhersage. [20, S. 4]

Anwendungsbereiche: Bilder, Texte;

Reichweite der Erklärungen: global, wenn viele Eingabewerte gestört werden oder globale Interpretationstechniken verwendet werden; lokal, wenn nur eine einzelne Zeile gestört wird oder wenn globale Interpretationstechniken eingesetzt werden; [8, S.34f]

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:
- die Methode kann leicht für ein Deep Neural Network implementiert werden;
- die Gradienten lassen sich mit relativ wenig Aufwand leicht über Backpropagation berechnen; [20, S. 4]

Nachteile:
- die Methode ist nicht optimal für aktuelle KI-Modelle, da verrauschte Gradienten (gradient shattering) und Erklärungsdiskontinuität auftreten; Varianten der Sensivity Analysis können diese Probleme teilweise lösen: SmoothGrad über Mittelwertbildung über den Gradienten und Integrated Gradient über die Bestimmung des Gradienten über einen bestimmten Pfad; [31, S. 13]
- die Berechnung der Methode über eine Optimierungsfunktion kann sehr rechenaufwändig sein; [31, S. 13]

Implementierung:
- SALib (Python);
- sensivity (R);

Integrated Gradient

Referenz: *Sundararajan et. al: Axiomatic Attribution for Deep Networks, 2017*

Modellbeschreibung: Die Methode verwendet zur Ermittlung der Erklärung die Annäherung an ein bestimmtes Integral. Dazu wird zunächst vom Originalbild ein Satz Kopien mit unterschiedlichen Helligkeitsstufen erstellt. Für diese Kopien werden die Gradienten berechnet. Über die Gradienten wird der Mittelwert gebildet und dann das elementweise Produkt des Mittelwerts mit dem Originalbild ermittelt. [37]

Die Autoren beschreiben Integrated Gradients als Zuordnungs-Methode (Attribution method), die alle dafür notwendigen Eigenschaften erfüllt: die Vollständigkeit, die Erhaltung der Linearität, die Symmetrieerhaltung und die Empfindlichkeit (Sensivity). [37]

Anwendungsbereiche: von den Autoren getestet für Bilder, Texte und chemische Modelle; [38, S. 1]

Die Methode kann als Verallgemeinerung von DeepLift betrachtet werden. [1, S. 182]

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:
- die Methode ist unabhängig von der Implementierung des Black-Box Modells;
- die Methode ist für jede Netzwerkarchitektur geeignet; [1, S. 182]

Nachteile:
- es besteht ein relativ hoher Rechenaufwand, da zur Bewertung des durchschnittlichen Gradienten ein Integral numerisch ermittelt werden muss. [1, S. 182]

Implementierung: IntegratedGradients (Python);[a]

[a] https://github.com/ankurtaly/Integrated-Gradients

6.2.2 Störungs-basierte Ansätze

Meaningful Perturbation

Referenz: *Fong, R. & Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation, 2017*

Modellbeschreibung: Meaningful Perturbation ist eine Methode die Erklärungen über Meta-Prädiktoren erzeugt. Es wird nach interpretierbaren Regeln gesucht, mit denen die Beziehung zwischen Eingabewerten und Ausgabe dargestellt werden kann. Um dieses Ziel zu erreichen werden minimale lokale Störungen für einzelne Datenpunkte erzeugt und analysiert, wie sich die Störungen auf die Vorhersage des Black-Box Modells auswirken.

Die Ergebnisse können in Saliency Maps dargestellt werden, in den für die Vorhersage relevante Regionen des Bildes hervorgehoben werden. [31, S. 146], [6, S. 1f]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Model-übergreifend;

Vorteile:
- die Störungen können flexibel in das Black-Box Modell eingebracht werden, dadurch können Eigenschaften der Erklärung gesteuert werden; [31, S. 146]
- über den Einsatz von Meta-Prädiktoren wird die Korrektheit von Erklärungen messbar; [6, S.2]

Nachteile:
- hoher Rechenaufwand;

Implementierung: Pytorch (Python); [a]

[a] https://github.com/ruthcfong/perturb_explanations

Prediction Difference Analysis

Referenz: *Zintgraf et al.: Visualizing Deep Neural Network Decisions: Prediction Difference Analysis, 2017*

Modellbeschreibung: Prediction Difference Analysis (PDA) bringt Störungen in die Pixelumgebung eines analysierten Merkmals ein und entfernt dadurch Informationen. Die Auswirkung dieses Vorgehens wird analysiert und auf deren Grundlage wird eine Bewertung der Relevanz der Merkmale vorgenommen. Um eine individuelle Klassifikationsentscheidung zu erklären, wird für jedes Merkmal ein Relevanzwert ermittelt, der den Beitrag des Merkmals zur Entscheidung für oder gegen eine vorhergesagte Klasse widerspiegelt. Die Ergebnisse werden in einer Saliency Heatmap dargestellt. [44, S. 1f]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: global;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:

- die Methode liefert Informationen darüber, welche Pixel des Bildes für und welche gegen eine bestimmte Klassifikation sprechen; [44, S. 8]

Nachteile:

- hoher Ressourcenaufwand für die Berechnung notwendig, der Einsatz von GPUs kann die Rechenzeit verkürzen; [44, S. 8]

Implementierung: DeepVis-PredDiff (Python);[a]

[a] https://github.com/lmzintgraf/DeepVis-PredDiff

Deconvolution / Occlusion

Referenz: *Zeiler & Fergus: Visualizing and Understanding Convolutional Networks, 2013*

Modellbeschreibung: Mit der Deconvolution-Methode können die für die Klassifikation eines Convolutional Neural Networks (CNN) bedeutsamen Bildbereiche ermittelt werden. Dies geschieht durch systematisches Abdecken von Teilbereichen des Eingabebildes mit einem grauen Rechteck. Nach der Abdeckung wird die Aktivierungsänderung ermittelt. Aus dem Ausmaß der Änderung kann die Bedeutung der verdeckten Bereiche für die Vorhersage ermittelt werden. [44, S. 2]

Anwendungsbereiche: Klassifizieren von Bildern;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:
- die Methode zeigt die Bedeutung von einzelnen Bildregionen für das Klassifikationsergebnis; [44, S. 2]

Nachteile:
- die Kanten des grauen Rechtecks, dass Abdeckung verwendet wird, können das Analyseergebnis beeinflussen, da CNNs sehr empfindlich auf Kanten reagieren;
- es kann passieren, dass Teile des Objekts, das identifiziert werden soll, verdeckt werden und dadurch ein falsches Vorhersageergebnis entsteht; [7]

Implementierung: DeconvNet (Python)[a]

[a] https://github.com/tdeboissiere/DeepLearningImplementations/tree/master/DeconvNet

SHAP/Shapley Values

Referenz: *Lundberg, Lee: A Unified Approach to Interpreting Model Predictions, 2017*

Modellbeschreibung: SHAP beruht auf Shapley-Werten, ein Konzept aus der kooperativen Spieltheorie. Das Ziel von Shapley-Werten ist, dass die Gewinne in einer Gruppe von mehreren Teilnehmern fair verteilt werden, wobei davon ausgegangen wird, dass die Gruppen-Teilnehmer zusammenarbeiten, die Beiträge der einzelnen Teilnehmer aber unterschiedlich sind. Der Shapley-Wert ist der durchschnittlich erwartete Beitrag eines Spielers, unter Berücksichtigung aller möglichen Gruppenkonstellationen.[a]

SHAP verwendet die Shapley-Werte um die Beiträge einzelner Merkmale für die Vorhersage des Black-Box Modells zu berechnen. Aus den Beiträgen werden die Gewichte für die Vorhersage-Funktion ermittelt. Je weniger Merkmale einer Gruppe zusammengefasst sind, desto genauer kann der Beitrag eines einzelnen Merkmals errechnet werden. [18, Kap. 5.10]

Anwendungsbereiche: Bildklassifikation, Tabellendaten;

Reichweite der Erklärungen: lokal und global durch die Aggregation von Shapley-Werten; [18, Kap. 5.10]

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:

- schnelle Implementierung für Baum-basierte Modelle;
- hohe Übereinstimmung zwischen globalen und lokalen Erklärungen der Black-Box, da die globale Modellinterpretation auf den lokalen Erklärungen beruht;
- die Methode basiert auf einer soliden theoretischen Grundlage (Spieltheorie); [18, Kap. 5.10]

Nachteile:

- die Methode ist extrem rechenaufwändig und daher langsam;
- durch den Fokus auf Korrelationen von Merkmalen, kann die Bedeutung einzelner Merkmale verzerrt werden;
- die Methode ermittelt die Bedeutung eines Merkmals für die Vorhersage. Sie macht aber keine Aussage darüber, wie sich eine Änderung eines Merkmal-Wertes auf die Vorhersage auswirkt; [b]

Implementierung:

- shap (Python)[c]
- shapper (R)[d]

[a] https://www.investopedia.com/terms/s/shapley-value.asp

[b]https://medium.com/civis-analytics/demystifying-black-box-models-with-shap-value-analysis-3e20b536fc80
[c]https://github.com/slundberg/shap
[d]https://modeloriented.github.io/shapper/

6.3 Optimierungs-basierte Ansätze

Deep Generator Networks

Referenz: *Nguyen et al.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016*

Modellbeschreibung: Um zu verstehen wie ein Neuronales Netz funktioniert, kann man analysieren, welche Erkenntnis jedes Neuron hervorgebracht hat. Mit Hilfe von der Methode Aktivierungsmaximierung (AM) können Eingabewerte (zum Beispiel für ein Bild) so erzeugt werden, dass sie einen hohen Aktivierungsgrad eines Neurons bewirken. Im Bereich der Klassifizierung von Bildern wird durch Aktivierungsmaximierung zunächst ein Zufallsbild erzeugt und anschließend über Backpropagation iterativ ermittelt, wie jedes Pixel verändert werden sollte, damit die Aktivierung eines Neurons erhöht wird. [22, S. 1f]

Kurz gesagt, werden über Aktivierungsmaximierung Prototypen erzeugt, die das trainierte Modell repräsentieren. Die Erstellung der Prototypen geschieht über die Suche nach Eingabewerten, die die angestrebte Ausgabe des Modells erzeugen können. [31, S. 13]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:
- die Methode kann genutzt werden, um aus Bildern eine Filmsequenz oder aus textlichen Beschreibungen ein Bild zu erzeugen; [22, S. 8]

Nachteile:
- die Berechnung über eine Optimierungsfunktion kann sehr rechenaufwändig sein; [31, S. 13]

Implementierung: Source Code um die wichtigsten Ergebnisse des Referenzwerkes zu reproduzieren; [a]

[a]https://github.com/Evolving-AI-Lab/synthesizing

Deep Visualization

Referenz: *Yosinski et al.: Deep Visualization, 2015*

Modellbeschreibung: Ziel der Methode ist zu verstehen, was ein Neuron in einem Deep Neural Network (DNN) bewirkt, dazu werden bei Deep Visualization verschiedene Verfahren miteinander kombiniert: Vorwärtsaktivierungswerte, Gradientenanstieg, dominierende Bilder, Deconv-Hervorhebungen und Rückwärtsdifferenzen. Dadurch kann die Reaktion jedes Neurons eines trainierten Netzes auf das Eingabebild, interaktiv dargestellt werden. [40, S. 4]

Ähnlich wie bei Deep Generator Networks wird in dieser Methode ein repräsentativer Prototyp für das Black-Box Modell erstellt. Dabei werden mittels Aktivierungsmaximierung die Eingabewerte ermittelt, die eine gute Annäherung an die gewünschte Antwort des Modells erzeugen. [31, S. 13]

Anwendungsbereiche: Klassifizierung von Bildern;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: geeignet zur Untersuchung von Deep Neural Networks;

Vorteile:

- die Methode ist robust gegenüber Änderungen von Parametern bezüglich Maßstab, Belichtung, Kontext und wie das Objekt positioniert ist; [40, S. 5]

Nachteile:

- der Einsatz einer Optimierungsfunktion kann zu einem hohen Rechenaufwand führen; [31, S. 13]

Implementierung: Deep Visualization Toolbox (Python)[a]

[a] https://github.com/yosinski/deep-visualization-toolbox

6.4 Ausbreitungs-basierte Ansätze

Layerwise Relevance Propagation (LRP)

Referenz: *Bach et al.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, 2015*

Modellbeschreibung: LRP verwendet die Rückwärtszerlegung der Vorhersage eines Neuronalen Netzes, um das Vorgehen des Modells zu erklären. Die Methode führt nach einem Standard-Vorwärtsdurchlauf einen spezifischen Rückwärtsdurchlauf durch, bei dem für jeden Layer des Netzwerks eine bestimmte Ausbreitungsregel (Propagation Rule) zur Anwendung kommt. Dadurch wird für jedes Neuron des Netzwerkes ein Relevanz-Wert ermittelt. [2, S. 3]

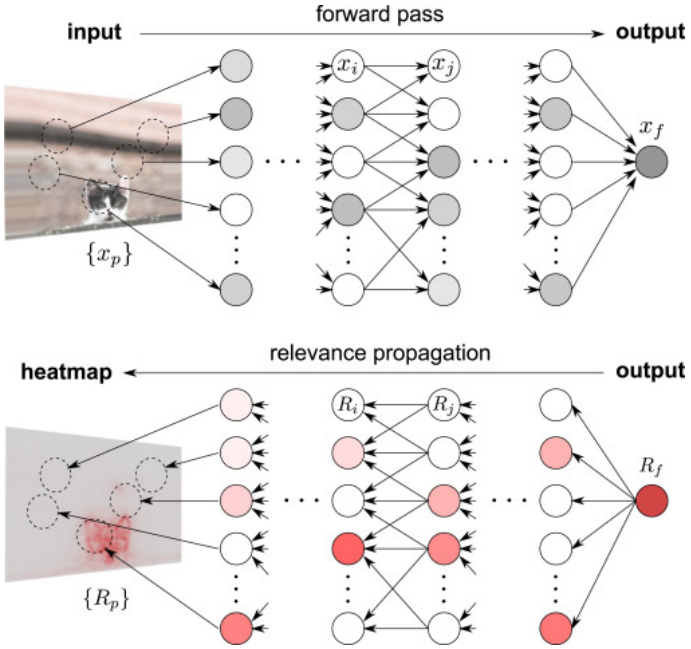


Abbildung 6.2: Backwardspropagation in LRP
(Quelle: http://danshiebler.com/2017-04-16-deep-taylor-lrp/)

Die Relevanz-Werte der einzelnen Merkmale für die Vorhersage wird in einer Heatmap visuell dargestellt. Aus der Heatmap wird ersichtlich, welchen Beitrag jedes Merkmal zum Klassifizierungsergebnis beigetragen hat. [12, S. 138]

Anwendungsbereiche: Bild- und Text-Klassifikation, Sprache, Videos, Gesichtserkennung, Spiele;

Reichweite der Erklärungen: global und lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für Neuronale Netze;

Vorteile:
- relativ geringer Rechenaufwand;
- Verzerrungen in Modellen und Datensätzen können erkannt werden; [3, S. 215]
- LRP kann für unterschiedliche Anwendungsfälle eingesetzt werden (Anwendungsbereiche siehe oben);

Nachteile:
- LRP erfüllt nicht das Kriterium der Implementation Invarianz und daher kann es vorkommen, dass unwichtige Aspekte des Modells überbewertet werden; [38, S. 2f]
- teilweise ist LRP numerisch instabil; [13]

Implementierung:
- iNNvestigate neural networks (Python);[a]
- TensorFlow LRP Wrapper (Python);[b]
- LRP Toolbox (Python);[c]

[a] https://github.com/albermax/innvestigate
[b] https://github.com/VigneshSrinivasan10/interprettensor
[c] https://github.com/sebastian-lapuschkin/lrp_toolbox

Deep Taylor Decomposition (DTD)

Referenz: *Montavon et al.: Explaining nonlinear classification decisions with deep Taylor decomposition, 2017*

Modellbeschreibung: Die Methode verwendet zur Erklärung der Vorhersage des Black-Box Modells die Eingabewerte(Pixel) des Modells. Im Rückwärtsdurchlauf durch das Neuronale Netz des Black-Box Modells wird der Ausgangswert auf die Eingangsvariablen zerlegt (decomposition). In dieser Methode wird die Funktion, die die Beziehung zwischen der Aktivierung von Layern und der Aktivierung von Knoten im Neuronalen Netz beschreibt in partielle Ableitungen zerlegt. Durch diese partiellen Ableitungen kann die Relevanz-Ausbreitungsfunktion des Black-Box Modells approximiert werden. Dazu werden Taylor-Reihen verwendet. Über eine Heatmap werden die für die Klassifizierung relevanten Bereiche des Ausgangsbildes hervorgehoben.[a,b]

Anwendungsbereiche: Klassifikation von Bildern;

Reichweite der Erklärungen: global und lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für Neuronale Netze;

Vorteile:

- unter verschiedenen Architekturen und Datensätzen verhält sich die Methode stabil, ohne dass die Hyperparameter dafür angepasst werden müssten; [21, S. 220]

Nachteile:
- keine bekannt;

Implementierung:

iNNvestigate neural networks (Python)[c]

[a] **http://www.heatmapping.org/deeptaylor/**.
[b] **http://danshiebler.com/2017-04-16-deep-taylor-lrp/**.
[c] **https://github.com/albermax/innvestigate**.

Deep Lift (Deep Learning Important FeaTures)

Referenz: *Shrikumar et al.: Learning Important Features Through Propagating Activation Differences, 2017*

Modellbeschreibung: Über Backpropagation zerlegt DeepLift die Klassifizierungsvorhersage auf die Eingabewerte. Dabei wird die Relevanz der Beiträge aller Neuronen des Netzwerkes ermittelt. DeepLift vergleicht die Aktivierung jedes Neurons mit einer Referenzaktivierung. Über die Differenz der beiden Werte wird der Beitrag jedes Neurons bestimmt. Die Auswahl der richtigen Referenzwerte ist ausschlaggebend für die Qualität der Ergebnisse. Für die Erzeugung passender Referenzwerte ist fachspezifisches Wissen notwendig. [33, S. 1ff]

Anwendungsbereiche: Bild- und Ziffern-Klassifikation; klassifizieren von DNA-Sequenzen;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für Neuronale Netze;

Vorteile:
- durch die Verwendung der Differenz zum Referenzwert kann die Propagation auch dann eingesetzt werden, wenn der Gradient den Wert Null hat; dies wäre bei Recurrent Neuronal Networks (RNN) hilfreich; [33, S. 8]
- durch DeepLift können Abhängigkeiten erkannt werden, die andere Methoden nicht bemerken. [33, S.1]

Nachteile:
- DeepLift erfüllt nicht das Kriterium der Implementation Invarianz. Es kann daher zu einer Überwertung von unwichtigen Aspekten des Modells kommen. [38, S. 2f]

- es ist unklar, wie gute Referenzwerte empirisch aus den Daten ermittelt werden können; [33, S. 9]
- bisher wurde nicht spezifiziert, wie DeepLift auf RNNs angewendet werden kann; [33, S. 9]

Implementierung: deeplift (Python) [a,b]

[a] https://github.com/kundajelab/deeplift
[b] https://pypi.org/project/deeplift/

Excitation Backpropagation

Referenz: *Zhang et al.: Top-down Neural Attention by Excitation Backprop, 2016*

Modellbeschreibung: Excitation Backpropagation ist ein Backpropagation Schema das von einem pyramidenförmigen neuronalen Netzwerk ausgeht. Bei dem Modell wird das Netz zunächst von unten nach oben durchlaufen, um die Stimuli des Netzes zu ermitteln. Anschließend werden in umgekehrter Richtung von oben nach unten die relevantesten Neuronen bestimmt. In Attention Maps (Heatmaps) werden die Beziehungen zwischen den Vorhersagewerten und den Eingangssignalen visuell dargestellt.

Das Verfahren beruht auf einem WTA-Schema (Winner-Takes-All), mit dem untersucht wird, welche Neuronen eines neuronalen Modells mit der höchsten Wahrscheinlichkeit für die Aktivierungen der höheren Ebene verantwortlich sind. [41, S. 1f]

Anwendungsbereiche: Bildklassifikation, Zuordnung von Text zu Bildbereichen;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: spezifisch für Convolutional Neural Network (CNN); [41, S. 1]

Vorteile:
- das Verfahren arbeitet genau und ist vielseitig einsetzbar;
- die Berechnung kann effizient in einem einzigen Rückwärtsdurchlauf durchgeführt werden; [41, S. 17]

Nachteile:
- keine bekannt;

Implementierung: Caffe-ExcitationBP (Python)[a]

[a] https://github.com/jimmie33/Caffe-ExcitationBP

Guided Backpropagation

Referenz: *Springenberg et al.: Striving for Simplicity: The All Convolutional Net, 2015*

Modellbeschreibung: Guided Backpropagation ermittelt nicht den Beitrag von einzelnen Eingabefeatures für die Vorhersage, sondern versucht Muster bei den Eingabewerten zu finden, die die Ausgabe des Netzwerkes bestimmt haben. Um dies zu erreichen, setzt Guided Backpropagation auf eine Veränderung der Architektur des neuronalen Netzes. Viele CNNs basieren auf abwechselnde Convolution- und Max-Pooling-Layer. Zur Optimierung der Bilderkennung ersetzt Guided Backpropagation die Max-Pooling Layer durch Convolutional Layer, ohne dass es dadurch zu einer Beeinträchtigung der Genauigkeit kommt. Das Verfahren basiert auf dem Deconvolutional Ansatz. Deconvolution wird hier zum Visualisieren der Konzepte, die durch die höheren Layer erlernt wurden, eingesetzt. [36, S. 1f und S.8]

Bei dem Verfahren werden die negativen Gradienten auf null gesetzt. Dadurch können im Backpropagation-Prozess aussagekräftigere Heatmaps erzeugt werden. [4, S. 5]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für CNNs;

Vorteile:
- dje Methode ist relativ einfach;
- die Methode erzeugt scharfe Abbildungen der fokussierten Bildbereiche; [36, S. 9]

Nachteile:
- die Methode ist anfällig für Störungen, besonders bei homogenen Hintergrund; [35, S. 5]
- die Methode zeigt die Anfälligkeit der Ausgabe wenn Änderungen bei den Eingabewerten vorgenommen werden, die Anfälligkeit ist aber nicht Grundlage der Vorhersage des Black-Box Modells; [4, S. 2]

Implementierung: verschiedene Bibliotheken für Python;[a]

[a] https://github.com/topics/guided-backpropagation

Grad-CAM (Gradient-weighted Class Activation Mapping)

Referenz: *Selvaraju et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016*

Modellbeschreibung: Grad-CAM basiert auf CAM (Class Activation Mapping). CAM ermittelt den von einem Convolutional Neural Net (CNN) verwendeten Bereich, der für eine Klassifizierung ausschlaggebend ist. Dabei versucht CAM durch Reduzierung der Komplexität des Modells einen höheren Grad an Erklärbarkeit zu erlangen. Grad-CAM verwendet die CAM-Methode, jedoch ohne die Komplexität des Modells zu verändern. Während CAMs nur für bestimmte CNN-Architekturen anwendbar ist, gilt dieser Einschränkung für Grad-CAMs nicht. [32, S. 2f]

Grad-CAM verwendet Gradienteninformationen aus der letzten Convolutional-Schicht eines CNN, um eine grobe Feature Map zu erstellen, in der die für die Vorhersage wichtigen Bereiche im Ausgangsbild hervorgehoben werden.[a]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für CNN-basierte Architekturen;

Vorteile:
- Grad-CAM-Visualisierungen haben eine hohe Wiedergabetreue; [32, S. 6]
- hoher Interpretierbarkeit; [32, S. 6]
- Klassen werden genau unterschieden;
- Verzerrung in Datensätzen werden durch die Methode erkannt; [32, S. 9]
- die Methode kann einzelne Objekte in einem Bild hervorheben (z.B. in einem Bild mit einem Hund und einer Katze wird eins der beiden Tiere fokussiert); [32, S. 2]

Nachteile:
- Grad-CAM kann Klassen im Bild unterscheiden und wichtige Bereiche hervorheben, hat aber Schwächen im feinkörnigen Pixel-Bereich;[b]
- die Methode ist nur für bestimmte Architekturen (die AveragePooling einsetzen) geeignet;
- die Erstellung einer gröberen Relevanz-Heatmap kann zu Artefakten und Informationsverlusten führen;[c]

Implementierung: grad-cam (Python);[d]

[a](http://gradcam.cloudcv.org/

[b]https://medium.com/@mohamedchetoui/grad-cam-gradient-weighted-class-activation-mapping-ffd72742243a

[c]http://blog.qure.ai/notes/deep-learning-visualization-gradient-based-methods

[d]https://github.com/ramprs/grad-cam

PatternNet und Pattern Attribution

Referenz: *Kindermans et al.: Learning how to explain neural networks: PatternNet and PatternAttribution, 2017*

Modellbeschreibung: PatternNet wird verwendet, um das „Signal“ eines Bildes sichtbar zu machen. Mit „Signal“ werden die relevanten Bereiche eines Bildes (z.B. ein Objekt) bezeichnet, die im Neuronalen Netz Aktivitäten ausgelöst haben. Es wird ermittelt, welche Eingangswerte die Aktivierungen, dargestellt in der Heatmap, ausgelöst haben. Die nicht-relevanten Teile und das Rauschen werden entfernt.

Mit Hilfe von PatternAttribution wird die Relevanz einzelner Pixel für die Klassifikationsentscheidung ermittelt. Die Zuordnung (attribution) gibt die Relevanz der jeweiligen Bereiche eines Bildes an. Zur Bestimmung der Zuordnung kann zum Beispiel der Ansatz der Backpropagation verwendet werden. [15, S. 1ff]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für Neuronale Netze;

Vorteile:
- das Verfahren erzeugt sehr gute Ergebnisse für lineare Modelle und gute für Bilder;
- die Berechnung der individuellen Erklärungen benötigt relativ wenig Rechenaufwand; [15, S. 8ff]

Nachteile:
- für die Trainingsphase, in der die Unterscheidung zwischen Signal und Rauschen getroffen wird, ist erhöhter Rechenaufwand notwendig; [15, S. 5ff]

Implementierung: im Paket iNNvestigate (Python)[a] enthalten;

[a]https://github.com/albermax/innvestigate

6.5 Erklärung des Modells

SpectralRelevanceAnalysis (SpRAy)

Referenz: *Lapuschkin et al.: Unmasking Clever Hans predictors and assessing what machines really learn, 2019*

Modellbeschreibung: SpRAy ist eine Weiterentwicklung von LRP und hat das Ziel das Entscheidungsverhalten von Black-Box Modellen und großen Datensätzen zu erkennen und zu bewerten. Zunächst werden Relevanzklassen für Datenstichproben und interessante Objektklassen mit Hilfe von LRP bestimmt. Die Ergebnisse, dargestellt in verschiedenen Heatmaps, werden dann mit einer Spektralanalyse geclustert. Jedes Cluster entspricht einer vom Modell gelernten Vorhersagestrategie. Aus den erzeugten Clustern lassen sich so, die Vorsagestrategien für ein Objekt ermitteln. [17, S.3ff]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: global;

Modell-spezifisch oder Modell-übergreifend: Modell-übergreifend;

Vorteile:
- mit der Methode können Schwachstellen in Modellen und Datensätzen erkannt werden; [31, S. 14]
- auch für große Datensätze geeignet;
- es können Unregelmäßigkeiten im Entscheidungsprozess eines Modells erkannt werden, zum Beispiel wenn das Erkennen eines Objekts nicht durch die Bildanalyse, sondern durch die Analyse der Metadaten zum Bild erfolgt. [17, S. 4]

Nachteile:
- keine bekannt;

Implementierung: bisher keine veröffentlicht;

Feature Visualization

Referenz: *Erhan et al.: Visualizing Higher-Layer Features of a Deep Network, 2009*

Modellbeschreibung: Feature Visualization zeigt schrittweise, wie ein Neuronales Netz über die verschiedenen Schichten des Netzes sein Verständnis über das fokussierte Bild aufbaut.

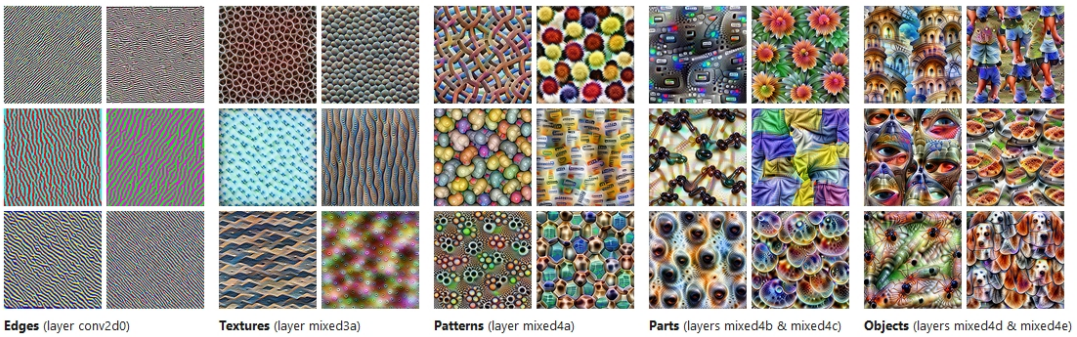


Abbildung 6.3: Feature Visualization auf den verschiedenen Layern [23]

Die Methode optimiert die Eingangsdaten mit Hilfe von Ableitungen iterativ auf ein Ziel hin. Es wird erkundet, welche Eingabedaten die Ursache für ein bestimmtes Verhalten (eine interne Neuronen Aktivierung oder die endgültige Ausgabe) sind. Der Optimierungsansatz hilft beim Verständnis, wonach ein Modell sucht, dies können zum Beispiel Neuronen, Kanäle, Layer oder Klassenwahrscheinlichkeiten sein. [23]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: global;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für NN;

Vorteile:

- Die Methode ermöglicht eine Einsicht über das schrittweise Vorgehen eines Neuronalen Netzes bei der Klassifizierung eines Bildes.
- Durch eine Kombination mit anderen Methoden kann ermittelt werden, welche Pixel für eine Klassifizierung ausschlaggebend waren. [18, Kap. 7.1.1]

Nachteile:

- Die Ergebnisse der Methode sind nicht immer für Menschen verständlich.
- Man kann der trügerischen Illusion unterliegen, dass man das Neuronale Netz durch das Ergebnis der Feature Visualization umfassend versteht. [18, Kap. 7.1.1]

Implementierung: tensorflow/lucid (Python)[a]

[a]https://github.com/tensorflow/lucid

Network Dissection

Referenz: *Zhou et al.: Interpreting Deep Visual Representations via Network Dissection, 2017*

Modellbeschreibung: Mit Network Dissection ist es möglich zu erkunden, was in den einzelnen Bereichen eines tiefen Convolutional Neuronal Networks (CNN) geschieht. Dazu wird die Ausgabe jedes Bereichs mit verschiedenen visuellen semantischen Konzepten abgeglichen. Jeder Schicht (layer) des CNNs werden Konzepte aus der realen Welt zugeordnet (zum Beispiel: Farbe, Material, Oberflächenstruktur, Teile, Objekte, Szenen), die am besten zu dieser Schicht passen. Durch die Methode können die verborgenen Bereiche von CNNs interpretierbar gemacht werden und verschiedene Netzwerkarchitekturen miteinander verglichen werden. [42, S. 1]

Die Autoren verstehen unter Interpretierbarkeit den Grad der Übereinstimmung von einer tiefen visuellen Repräsentation mit für Menschen verständlichen Konzepten. Die Methode erzeugt Interpretierbarkeit in drei Schritten:

1. Datenerfassung: Mit Hilfe eines Vergleichs-Datensatzes (Broden-Datensatz[a]) wird eine große Zahl der für Menschen verständlichen Konzepte im Ausgangsdatensatz identifiziert.

2. Netzwerkaktivitäten ableiten: Es wird geprüft, wie die versteckten Variablen (layer) des CNN auf die verschiedenen Konzepte reagieren, zum Beispiel wird ermittelt, ob ein Schwellenwert überschritten wurde.

3. Abgleich zwischen Aktivitäten und Konzepten: Der Grad der Übereinstimmung zwischen versteckten Variablen und den Konzepten wird berechnet. [42, S. 2f]

Anwendungsbereiche: Bilderkennung;

Reichweite der Erklärungen: lokal;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für CNNs;

Vorteile:
- die Methode ermöglicht neue Erkenntnisse über den hierarchischen Aufbau von CNNs; [42, S. 12]

Nachteile:
- es können nur die Konzepte erkannt werden, die im verwendeten Datensatz (hier: Broden) enthalten sind;

- Konzepte, die schwierig zu benennen sind (z.B. Ecke eines Raums), werden nicht erkannt;
- wenn ein Layer ein feinkörniges Konzept auswählt (zum Beispiel: Stuhlbein aus Holz), könnte ein gröberes Konzept (zum Beispiel: Stuhl) mit einen geringeren Übereinstimmungsgrad bewertet werden; [42, S. 12]
- Es werden Datensätze benötigt, in denen jedes Pixel mit einem konzeptuellen Label benannt wurde. [18, Kap. 7.1]

Implementierung: NetDissect (Python);[b]

[a]Der Broden-Datensatz enthält 60.000 Bilder, mit 1.000 visuellen Konzepten. [18, Kap. 7.1]

[b]https://github.com/CSAILVision/NetDissect

Testing with Concept Activation Vectors (TCAV)

Referenz: *Kim et al.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), 2017*

Modellbeschreibung: Das Modell beruht auf Concept Activation Vector (CAV), welches die Interpretation des internen Zustands eines Neuronalen Netzes mit menschlichen Konzepten (z.B. Farbe, Form etc.) ermöglicht. CAV zeigt den Bedeutungswert eines bestimmten Konzeptes für die Vorhersage des Modells. TCAV bewertet die jeweilige Bedeutung von vielen verschiedenen Konzepten, für das konkret Ausgangsbild. Die vom Netzwerk auf den einzelnen Layern erlernten Konzepte werden mit menschlichen Konzepten verglichen und ihre Übereinstimmung bewertet, zum Beispiel wird das menschliche Konzept „Zebra“ mit dem vom Netz erlernten Konzept „Streifen“ verglichen. Da diese beiden Konzepte eine hohe Übereinstimmung aufweisen, wird der Wahrscheinlichkeitswert für die Interpretation des Bildes als „Zebra“ erhöht.

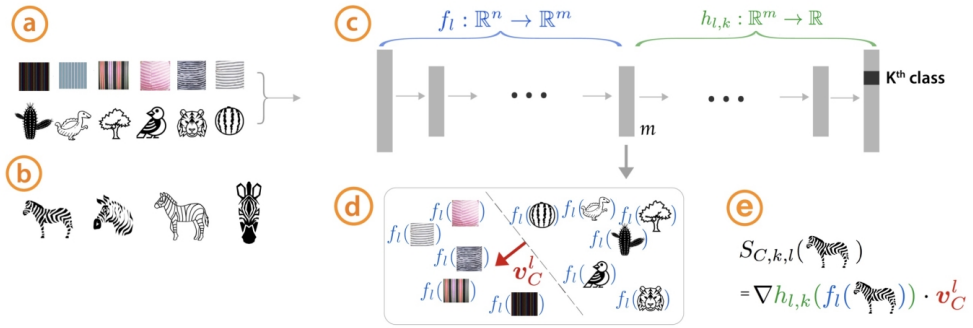


Abbildung 6.4: TCVA Beipiel: erkennen eines Zebra [14, S. 2]

Die verschiedenen Konzepte werden im Neuronalen Netz als Vektoren dargestellt und die Wahrscheinlichkeit der Modellvorhersage wird über Richtungsableitungen quantifiziert. [14, S. 1ff]

Anwendungsbereiche: Klassifizierung von Bildern, Audio, Videos, Sequenzen; [14, S. 8]

Reichweite der Erklärungen: global;

Modell-spezifisch oder Modell-übergreifend: Modell-spezifisch für Neuronale Netze;

Vorteile:

- eine für Nicht-Experten verständliche Erklärung des internen Zustand eines Deep Learning Modells wird erzeugt;
- die Modell-Entscheidungen können mit natürlichen intuitiven Konzepten erklärt werden; die Konzepte müssen dafür nicht vor der Trainingsphase bekannt sein, sondern können im Nachhinein angegeben und konkretisiert werden; [14, S. 8]

Nachteile:

- keine bekannt;

Implementierung: tensorflow/tcav; [a]

[a] https://github.com/tensorflow/tcav