Experiment zur Evaluierung der Nützlichkeit von Interpretationsmethoden für neuronale Netze

Juri Zach
Hochschule für Angewandte Wissenschaften Hamburg
Berliner Tor 5, 20099 Hamburg
Deutschland
juri.zach@haw-hamburg.de

***Zusammenfassung*—Dieser Artikel befasst sich mit dem praktischen Nutzen von Interpretationsmethoden für neuronale Netze, auf einer theoretischen Ebene. Es werden verschiedene Interpretationsmethoden vorgestellt und die Schwierigkeiten beim Testen derselben erläutert. Anschließend wird ein alternativer Ansatz vorgeschlagen, mit dessen Hilfe der praktische Nutzen von Interpretationsmethoden experimentell belegt werden kann. Dies wird mithilfe von mehreren Beispielen veranschaulicht.**
Das Ziel dieses Artikels ist es, den praktischen Einsatz von Interpretationsmethoden zu fördern um einen sicheren und ethischen Einsatz von künstlicher Intelligenz zu ermöglichen.

***Index Terms*—neuronale Netze, Interpretierbarkeit, maschinelles Sehen, Faltungsnetze**

I. Einleitung

Die Forschung im Bereich der künstlichen Intelligenz hat in den letzten Jahren große Fortschritte gemacht und findet sich zunehmend im alltäglichen Leben wieder. Bekannte Beispiele sind Facebook-Newsfeeds, Chatbots, Spracherkennung, digitales Marketing oder Einparkhilfen im Auto. Doch auch sicherheitskritische Anwendung wie beispielsweise intelligente Roboter, medizinische Diagnosesysteme oder autonom fahrende Autos werden stark vorangetrieben. Für viele dieser Aufgaben werden tiefe neuronale Netze eingesetzt. Diese sind in der Lage, die hoch dimensionalen Funktionen zu approximieren, mit denen komplexe Aufgaben wie Bild- und Spracherkennung, oder das Steuern von komplexen Roboter Aktoren, beschrieben werden.
Allerdings hat diese Technologie gerade im Bereich der Sicherheit und Testbarkeit starke Schwachstellen. Aufgrund ihrer Komplexität ist es kaum möglich nachzuvollziehen was ein neuronales Netz gelernt hat und wie es seine Entscheidungen trifft.
Der Forschungsbereich der erklärbaren künstlichen Intelligenz befasst sich unter anderem mit der Interpretierbarkeit von neuronalen Netzen. Auch hier wurden in den letzten Jahren große Fortschritte gemacht. Doch im Gegensatz zu den neuronalen Netzen bleibt dieser Forschungsbereich theoretisch und findet kaum praktische Anwendung.

Um den Einsatz der Interpretation von neuronalen Netzen voranzutreiben, stellt dieser Artikel vielversprechende Interpretationsmethoden vor und beschreibt Experimente mit dem ihr praktischer Nutzen evaluiert werden kann.
Die vorgestellten Interpretationsmethoden und Experimente können in verschiedenen Domänen eingesetzt werden, doch der Einfachheit halber beschränkt sich dieser Artikel ausschließlich auf das Anwendungsfeld des maschinellen Sehens.

Der Abschnitt II verweist auf verschiedenen wissenschaftliche Arbeiten, auf denen dieser Artikel aufgebaut ist. Im Abschnitt III wird beschrieben, wie der praktische Nutzen von Interpretationsverfahren ermittelt werden kann. In den Abschnitten IV bis VI werden exemplarische Experimente zu den verschiedenen Interpretationsmethoden beschrieben. Zum Abschluss werden im Abschnitt VII weitere Forschungen erläutert, mit denen die Erfolgschancen der Experimente maximiert werden können.

II. Ähnliche Arbeiten

Dieser Artikel baut auf verschiedenen wissenschaftlichen Arbeiten auf, welche sich sowohl mit theoretischen Aspekten als auch mit konkreten Implementierungen und Untersuchungen von Interpretationsmethoden beschäftigen.
Die wissenschaftlichen Arbeiten der Autoren Doshi-Velez und Kim [13] und Lipton [14] befassen sich mit den theoretischen Aspekten der Interpretierbarkeit von maschinellen Lernmethoden und beschreiben unter anderen, welche Anforderungen diese erfüllen sollten. Nach Lipton [14] entsteht die Notwendigkeit der Interpretierbarkeit, sobald ein neuronales Netz auf Werte optimiert werden soll, welche sich nicht als mathematische Funktion umsetzen und optimieren lassen. Die Aufgabe der Interpretationsmethode ist es sicherzustellen, dass diese Werte umgesetzt werden.
Beispiele hierfür sind:

- **Sicherheit:** Für komplexe Deep Learning Anwendungen ist es nicht praktikabel alle möglichen Anwendungsszenarien zu Testen. Interpretationsmethoden sollen es ermöglichen Sicherheitslücken gezielt zu identifizieren.
- **Wissenschaftlicher Erkenntnissgewinn:** Obwohl neuronale Netze nur darauf trainiert werden Annahmen zu treffen, besteht die Hoffnung aus ihrem gelernten Wissen und Entscheidungen, Erkenntnisse über die reale Welt zu ziehen. Hierfür ist jedoch ein Einblick in die *Blackbox* vonnöten.

- **Ethik:** Sowohl Politiker und Journalisten als auch Wissenschaftler fordern, dass die von künstlicher Intelligenz getroffenen Entscheidungen ethischen Standards entsprechen. [16]
- **Vertrauen:** Um das Vertrauen des Benutzers herzustellen, ist es vorteilhaft die Vorhersage des neuronalen Netzes durch Begründungen zu erweitern. Hierdurch können die Stärken und Schwächen des neuronalen Netzes abgeschätzt werden.

Zusätzlich zu den theoretischen Überlegungen finden sich in der Literatur viele Vorschläge zur praktischen Umsetzung der Interpretation von neuronalen Netzen. Die meisten wissenschaftlichen Arbeiten fokussieren sich auf Feature Visualisierung [6] [4] [15] [7] und Attribution [7] [8] [9] [10]. Zusätzlich hat die Mensch-Computer-Interaktions-Gemeinde, an vielversprechenden Interfaces geforscht [11] [12].

In neueren Arbeiten wurde die Aussagekraft der Interpretationen durch Kombination verschiedener Interpretationsmethoden und deren Erweiterung mit wirkungsvollen Interfaces, erneut gesteigert. Dazu gehört unter anderen die Arbeit von Olah et al. [1], in der verschiedene Varianten der Feature Visualisierung vorgestellt werden und gezeigt wird wie diese mit Attributionsmethoden und wirkungsvollen Interfaces kombiniert werden können. Die Arbeit von Carter et al. [2] zeigt, wie die vom neuronalen Netzen gelernten Konzepte als Aktivierungsvektor definiert und visualisiert werden. In eine Arbeit anonymer Autoren [17] werden von Menschen definierte Konzeptaktivierungsvektoren dazu genutzt, die Entscheidungen von neuronalen Netzen zu erklären.
Obwohl diese Entwicklungen sehr vielversprechend wirken, gibt es nach Recherchen des Autors noch keine wissenschaftliche Arbeit darüber, die evaluiert, wie gut sich heutige Interpretationsmethoden dazu eignen die theoretischen Anforderungen zu erfüllen, welche von Wissenschaftlern wie Doshi-Velez, Kim oder Lipton gefordert wurden.
Um dazu beizutragen diese Forschungslücke zu schließen wird im folgendem Abschnitt eine Methode vorgeschlagen, mit der empirisch belegt werden kann, ob eine Interpretationsmethode die theoretischen Anforderungen erfüllt und sich für die praktische Anwendung eignet.

III. Genereller Experiment Aufbau zum Evaluieren der Nützlichkeit von Interpretationsmethoden

Ein neuronales Netz wird oft als Black Box bezeichnet, obwohl die Eingabewerte, alle Gewichte und Rechenschritte bekannt sind. Diese sind allerdings zu viele und zu komplex vernetzt, als dass der Mensch sie in ihrer Gesamtheit begreifen könnte. Interpretationsmethoden versuchen eine Teilmenge dieser Komplexität auf ein, für Menschen erfassbares Maß zu reduzieren. Die hieraus resultierende Information muss anschließend, in einer für den Menschen verständlichen Domäne dargestellt werden. Für gewöhnlich wird hierfür die Domäne der Eingangswerte verwendet, da diese eine semantische Bedeutung haben.

Abbildung 1. Beispiel einer Feature Visualisierung aus einem mit ImageNet trainierten InceptionV1 Netzwerks. Diese Bild würde durch die Aktivierungs-Maximierung eines einzelnen Filters aus der Schicht *mixed5a* hergestellt.

Durch die Reduktion der Komplexität und das Überführen in eine andere Domäne, muss das Resultat der Interpretationsmethode nun wiederum von einem Menschen interpretiert werden, um Schlüsse über das neuronale Netz oder den Trainingsdatensatz zu ziehen.

Das folgende Beispiel erklärt diesen Vorgang anhand der Feature Visualisierung im Bereich des maschinellen Sehens: Features werden innerhalb des neuronalen Netzes als abstrakte Vektoren dargestellt. Mit Hilfe der Feature Visualisierung kann ein einzelnes oder eine Teilmenge der Features (Reduktion der Komplexität) auf der Domäne der Eingangswerte, also als Bild, dargestellt werden (Überführung in eine für Menschen verständliche Domäne). Ein Beispiel Resultat ist in der Abbildung 1 zu sehen. Um aus diesem Bild Schlüsse über das neuronale Netz oder den Trainingsdatensatz zu ziehen, müssen die einzelnen Pixel vom Menschen als semantische Konzepte[1] erkannt und im Zusammenhang mit der Aufgabe und etwaigen Eigenschaften des neuronalen Netzes, interpretiert werden.

Da die menschliche Interpretation, bei allem dem Autor bekannten Interpretationsmethoden, eine Notwendigkeit für den praktischen Einsatz derselbe ist, kann die Aussagekraft und Qualität der Interpretationsmethode nicht direkt gemessen werden. Dieses Problem zeigt sich vor allem bei Interpretationsmethoden, welche Informationen aus den Tiefen schichten der neuronalen Netze darstellen [4] [1] [17] [15]. Doch gerade diese Methoden sind nach Meinung des Autors und anderer Wissenschaftler [4] [1] besonders aussagekräftig. Um dennoch eine Aussage über die Güte verschiedener Interpretationsmethoden treffen zu können, wird in diesem Artikel ein experimenteller Aufbau vorgeschlagen, mit dem der praktische Nutzen von Interpretationsmethoden empirisch belegt werden kann.

[1]Definition Konzept: Als Konzept wird in dieser Arbeit als eine abstrakte Idee eines Objektes definiert. Beispiel Konzepte aus dem Bereich des (maschinellen) Sehens sind: Augen, Streifen, Fell, Licht und viele mehr.

Definition: Eine Interpretationsmethode gilt genau dann als Nützlich, wenn aus ihren Resultaten korrekte Annahmen über das neuronale Netz oder dem Trainingsdatensatz getroffen werden können, welche eine oder mehrere der theoretischen Anforderungen an Interpretationsmethoden erfüllt (siehe Abschnitt II).

Um ein geeignete Interpretationsmethode auf ihren praktischen Nutzen zu überprüfen wird eine Hypothese über ihre Nützlichkeit aufgestellt, welche in einem Experiment belegt oder widerlegt werden kann. Diese Hypothese beinhaltet den Nutzen der Interpretationsmethode und die Rahmenbedingen in denen sie angewendet werden kann.

Da die Hypothese, aufgrund der menschlichen Interpretationskomponente, nicht direkt bewiesen werden kann, muss sie durch ein statistisches Verfahren belegt oder widerlegt werden. Hierfür wird eine Reihe von Datenpunkten erhoben, die eine Aussage über die Nützlichkeit der Interpretationsmethode geben. Ein einzelner Datenpunkt wird durch das Ausführen der folgenden Schritte erzeugt, welche in einem Experiment mehrfach durchgeführt werden:

1) Die zu untersuchende Interpretationsmethode wird auf ein geeignetes neuronales Netz angewendet.
2) Mithilfe menschlicher Interpretation wird aus dem Resultat, eine Annahme über das neuronale Netz oder dem Trainingsdatensatz aufgestellt, welche dem in der Hypothese definierten Nutzenanforderungen entspricht.
3) Zum Bestätigen der Annahme wird ein Testdatensatz erstellt, welcher die Annahme in seinen Daten widerspiegelt. Um die Signifikanz der Ergebnisse des Testdatensatzes zu ermitteln, wird zusätzlich noch ein Referenzdatensatz erstellt.
4) Mithilfe der beiden Datensätze und statistischer Verfahren wird die Annahme über das neuronale Netz oder den Trainingsdatensatz belegt oder widerlegt und somit ein Datenpunkt zu belegen oder widerlegen der Hypothese erzeugt.

In den folgenden Abschnitten IV, V und VI werden konkrete Durchführungen des soeben beschriebenen Experiments exemplarisch erläutert.

IV. Experiment 1: Visualisierung von Klassen

Die Interpretationsmethoden der Feature Visualisierung liefern Eingangswerte des neuronalen Netzes, welche ein bestimmtes Neuron besonders stark aktivieren. Im Bereich des maschinellen Sehens sind dies Bilder, welche die semantische Bedeutung des Neurons widerspiegeln sollen. Da Neuronen nicht isoliert arbeiten, sondern Konzepte durch eine Kombination mehrerer Neuronen signalisieren [2], sind Feature Visualisierungen von einzelnen Neuronen semantisch schwer zu deuten. Eine Ausnahme sind die Ausgangsneuronen eines Klassifikator Netzwerks, da diese jeweils eine einzelne Klasse repräsentieren.
Durch das Visualisieren eines Ausgangsneuron kann die dazugehörige Klasse, aus der Sicht des neuronalen Netzes dargestellt werden. Dies ermöglicht es Aussagen über die Konzepte zu formulieren, an denen das neuronale Netze die Klasse erkennt.
Hierdurch lassen sich Fehler im neuronalen Netz finden und mögliche Sicherheitslücken schließen.

A. Formulierung der Hypothese

Das Visualisieren der Ausgangsneuronen eines Klassifikator Netzwerks hat folgenden Nutzen:

1) Es können gezielt Fehler und somit etwaige Sicherheitslücken identifiziert werden.
2) Stärken und Schwächen des neuronalen Netzes können abgeschätzt werden, wodurch das Vertrauen des Benutzers gestärkt wird. [2]

B. Experimentaufbau

Für dieses Experiment wird ein vortrainiertes Klassifikator-Netzwerk benötigt. Es ist von Vorteil ein Multi-Klassifikator zu nehmen, da nicht klar ist, ob aus allen Klassenvisualisierungen Annahmen über das neuronale Netze oder den Trainingsdatensatz aufgestellt werden können.
Des weiterem wird eine Methode zur Feature Visualisierung benötigt. Die Empfehlung des Autors ist das Visualisieren durch Lernverfahren [4], da diese semantisch vielversprechend aussehen und ihre Informationen ausschließlich aus dem zu interpretierenden neuronalen Netz ziehen.
Um die Resultate der Interpretationsmethode zu interpretieren, wird ein Mensch benötigt. Dieser sollte Experte für die Klassifizierungsaufgabe des neuronalen Netzes sein, da es ihn sonst schwerfallen wird Fehler, Stärken oder Schwächen aufzudecken.
Zum Überprüfen der menschlichen Annahmen werden zudem Test und Referenzdaten benötigt. Da die Art der benötigten Daten von den Annahmen abhängt, können diese nicht im Vorfeld gesammelt werden.

C. Experimentdurchführung

Für die Durchführung des Experiments wird die Feature Visualisierungsmethode auf die Ausgangsneuronen des neuronalen Netzes angewendet. Hierdurch wird für jede Klasse eine Visualisierung aufgestellt, welche vom menschlichen Experten interpretiert werden kann.
Abhängig vom der Visualisierung und der menschlichen Interpretation, werden nun Annahmen aufgestellt. Die folgenden Abschnitte beschreiben erwartete Szenarien, die Annahmen welche daraus gezogen werden können und wie diese mithilfe von Test und Referenzdatensätzen bestätigt werden.

1) Szenario 1:

[2] Ob das Wissen um Stärken und Schwächen eines neuronalen Netzes wirklich das Vertrauen des Benutzers stärkt, ist eine Annahme und sollte in einer psychologischen Studie erforscht werden.

a) Beschreibung: Die Visualisierung einer Zielklasse beinhaltet Irrtümliche Objekte, welche nicht zur Zielklasse, sondern zu anderen Klassen oder Hintergrund Objekten gehören.

b) Formulieren der Annahme: Beispiele, welche die Irrtümlichen Objekte beinhalten, tragen zur Klassifizierung der Zielklasse bei, auch wenn diese nicht im Beispiel enthalten ist.

c) Bestätigen der Annahme: Um diese Annahme zu bestätigen wir ein Testdatensatz erstellt, welcher Beispiele der Irrtümlichen Objekte, nicht aber die Zielklasse enthält. Zudem wird ein Referenzdatensatz erstellt, welche weder die Irrtümlichen Objekte noch die Zielklasse enthält. Durch das Klassifizierten beider Datensätze kann ermittelt werden, ob die Vorhersagen für die Zielklasse beim Testdatensatz signifikant höher sind als die vom Referenzdatensatz und somit die Annahme bestätigt oder widerlegt werden.

2) Szenario 2:

a) Beschreibung: Die Visualisierung zeigt sehr aussagekräftige Konzepte der Klasse, bildet allerdings keine Komponenten ab, mit denen die Klasse nur schwer zu identifizieren ist.

b) Formulieren der Annahme: Das neuronale Netz hat zwar gelernt die Zielklasse zu identifizieren, tut dies allerdings nur an einigen wenigen und möglicherweise einfachen Konzepten. Schwere Beispiele, in denen diese Konzepte nicht vorkommen, können somit nicht richtig klassifiziert werden.

c) Bestätigen der Annahme: Um diese Annahme zu bestätigen, wird eine Testdatensatz erstellt, in dem die Zielklasse abgebildet ist, nicht aber die Konzepte an dem das neuronale Netz diese erkennt. Zusätzlich wird ein Referenzdatensatz erstellt, in dem die Zielklasse, inklusive der einfach zu erkennenden Konzepte, abgebildet ist. Durch das Klassifizierten beider Datensätze kann ermittelt werden, ob die Vorhersagen für die Zielklasse beim Testdatensatz signifikant niedriger sind als die vom Referenzdatensatz und somit die Annahme bestätigt oder widerlegt werden.

V. Experiment 2: Darstellen von ausschlaggebenden Konzepten

Eine häufige Frage beim Umgang mit neuronalen Netzen ist, wie sie ihre Entscheidungen treffen.
Um diese Frage zu beantworten, können Attributionsmethoden [7] [8] [9] [10] verwendet werden, welche die Zusammenhänge von Neuronen darstellen. Überwiegend werden mit diesen Methoden Bedeutungsmasken (Saliency Maps) erzeugt, um die Relevanz der einzelnen Eingangswerte (welche im Bereich des maschinellen Sehens, Pixeln entsprechen) für die Entscheidung des neuronalen Netzes darzustellen.
Eine Kritik von Olah et al. [1] zu dieser Herangehensweise ist, dass die Bedeutung eines Pixels stark mit der anderer Pixeln verwoben ist und somit keine individuelle Relevanz besitzt. Außerdem sind Pixel semantisch sehr weit von den abstrakten Konzepten entfernt, welche tiefe neuronale Netze für gewöhnlich vorhersagen.
Das wirkliche Potential der Attributionsmethoden ergibt sich, wenn diese in Kombination mit Feature Visualisierung, auf tiefen Faltungsschichten eines neuronalen Netzes angewendet werden.

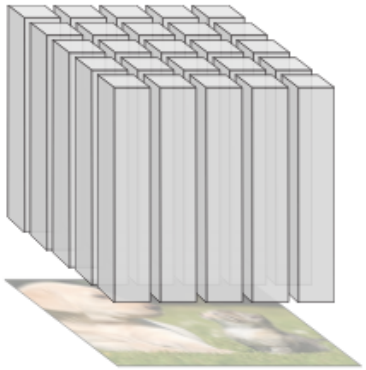

Abbildung 2. Die Aktivierung, aller Filter der gleichen räumlichen Lage einer Faltungsschicht, ist aus dem englischen auch als *Spartial-Activation* bekannt [1]. Die Rechtecke stellen die kombinierten Filter da, welche alle Informationen einer einzelnen räumlichen Lage eines Bildes beinhalten.

Um den neuronalen Aktivierungen der tiefen Schichten eine semantische Bedeutung zuzuweisen, wird eine Feature Visualisierungsmethode auf alle Filter-Aktivierungen der gleichen räumlichen Lage angewendet (siehe Abbildung 2). Hierdurch wird jeder Aktivierung der Ausgangsmatrix und somit jeder Position des Bildes ein abstraktes Konzept zugewiesen. Die Relevanz der verschiedenen Konzepte kann, mithilfe einer Attributionsmethode, ermittelt werden.
Dies ermöglicht es, die Entscheidung des neuronalen Netzes, anhand der Relevanz verschiedener semantischer Konzepte, zu argumentieren.

A. Formulierung der Hypothese

Das Visualisieren der Relevanz von tiefen Konzepten, für konkrete Beispiele, ermöglicht es:

1) Fehler in der Kausalität oder Interpretation des neuronalen Netzes zu entdecken und somit etwaige Sicherheitslücken gezielt zu identifizieren.
2) Anhand der Kausalität des neuronalen Netzes, wissenschaftliche Erkenntnisse zu gewinnen.
3) Die Entscheidungen eines neuronalen Netzes zu begründen und somit das Vertrauen des Benutzers zu festigen.

B. Experimentaufbau

Zur Durchführung dieses Experiments wird ein vortrainiertes Faltungsnetz benötigt. Zum Visualisieren der Faltungsschichten wird eine räumlich begrenzte Variante der Feature-Inversion-Methode verwendet, welche in Olah et al. [1] vorgestellt wird. Zusätzlich wird eine geeignete[3] Attributionsmethode benötigt, um die Relevanz der visualisierten Konzepte

[3]Hierfür eignen sich verschiedene Attributionsmethoden. Welche in diesem Kontext am besten funktioniert, ist noch nicht bekannt.

zu errechnen.
Da diese Interpretationsmethode nur für konkrete Beispiele funktioniert, sollten im Vorfeld einige interessante Datenpunkte ausgewählt werden. Um die Resultate der Interpretationsmethode zu interpretieren, wird ein Mensch benötigt. Dieser sollte Experte für die Aufgabe des neuronalen Netzes sein, da es ihn sonst schwerfallen wird die wirkliche Relevanz der Konzepte einzuschätzen.
Abhängig von den Annahmen des menschlichen Interpreters sind weitere Testdaten, sowie ein Referenzbild vonnöten, welches die Entscheidungen des neuronalen Netzes minimal beeinflusst. Dieses Bild kann durch Optimierungsverfahren ermittelt werden. Die Verwendung der Testdaten und des Referenzbildes wird im Kontext der erwarteten Szenarien beschrieben.

C. Experimentdurchführung

Für die Durchführung des Experiments wird das neuronale Netz auf einen einzelnen Datenpunkt angewendet. Hierbei werden die Aktivierungen der Schichten gespeichert. Mithilfe der Visualisierungs- und Attributionsmethode werden die Aktivierungen der Faltungsschichten visualisiert und deren Relevanz für die Entscheidung des neuronalen Netzes errechnet. Nun können die verschiedenen Schichten durch den menschlichen Experten interpretiert werden.
Die folgenden Abschnitte beschreiben erwartete Szenarien, die Annahmen, welche aus diesen gezogen werden können und wie diese bestätigt werden.

1) Szenario 1:

a) Beschreibung: In den Visualisierungen einer oder mehreren Faltungsschichten kann ein semantisches Konzept erkannt werden, welches stark zu einer bestimmten Entscheidung beiträgt. Dies kann sowohl ein Konzept sein, welches korrekterweise zur richtigen Entscheidung beiträgt, als auch ein Konzept, welches irrtümlicherweise zur falschen Entscheidung beiträgt.

b) Formulieren der Annahme: Das neuronale Netz hat das besagte Konzept gelernt und verwendet es für eine kausale Schlussfolgerung, zugunsten der besagten Entscheidung.

c) Bestätigen der Annahme: Um die Relevanz des Konzeptes zu bestätigen wird ein Testbild erzeugt. Hierfür wird das besagte Konzept aus dem Beispielbild herausgeschnitten und die entstehende Lücke durch das Referenzbild (Bild mit minimaler Relevanz) gefüllt. Für das Testbild wird die Vorhersage des neuronalen Netzes errechnet. Sollte sich das Ergebnis der Vorhersage signifikant zu Ungunsten der besagten Entscheidung verändern, belegt dies die Annahme, dass das Konzept stark zu der besagten Entscheidung beiträgt. Dieser Test bestätigt das Funktionieren der Attributionsmethode und kann auch in den folgenden Szenarien verwendet werden.
Um die Kausalität zwischen dem Konzept und der Entscheidung zu belegen, werden weitere Beispiele herausgesucht. Diese müssen das besagte Konzept beinhalten und zur selben Entscheidung führen. Sofern das Konzept auch bei den weiteren Testbeispielen, zur besagten Entscheidung beiträgt, belegt dies die Annahme über die Kausalität.

2) Szenario 2:

a) Beschreibung: In den Visualisierungen einer oder mehrerer Schichten, kann ein aus menschlicher Sicht relevantes Konzept erkannt werden, welches jedoch nicht zur Entscheidung des neuronalen Netzes beiträgt.

b) Formulieren der Annahme: Das neuronale Netz hat das besagte Konzept gelernt, doch die kausale Schlussfolgerung ist falsch, wodurch dem Konzept keine Bedeutung beigemessen wird.

c) Bestätigen der Annahme: Das Bestätigen der Annahme wird analog zum Szenario 1 durchgeführt. Der Unterschied besteht darin, dass nicht die Relevanz des Konzeptes, sondern dessen Irrelevanz belegt wird.

3) Szenario 3:

a) Beschreibung: Im Eingangsbild ist ein als relevant bewertetes Konzept zu sehen, welches aber zur falschen Entscheidung beiträgt. In späteren Schichten kann jedoch erkannt werden, dass das besagte Konzept vom neuronalen Netz, falsch interpretiert wird.
Ein Beispiel hierfür ist das Bild eines Schmetterlings dessen Flügel in der Eingangsschicht als relevant bewertet werden. Die Muster auf den Flügeln wurden in späteren Schichten (wie von der Natur beabsichtigt) jedoch als Gesicht eines Tieres interpretiert und tragen dadurch zu einer falschen Entscheidung bei.

b) Formulieren der Annahme: Der Unterschied zweier ähnlicher Konzepte wurde nicht ausreichend trainiert, wodurch diese leicht verwechselt werden.

c) Bestätigen der Annahme: Um diese Annahme zu überprüfen können weitere Beispiele herausgesucht werden, auf dem das falsch interpretierte Konzept abgebildet ist. Sollten diese ebenfalls zur besagten falschen Entscheidung beitragen, betätigt dies die Annahme. Zusätzlich kann mithilfe der räumlichen Feature-Inversion [1], die Interpretationen innerhalb der tiefen Schichten des neuronalen Netzes überprüft werden.

VI. Experiment 3: Bewertung der Relevanz ausgewählter Konzepte

Sobald neuronale Netze Entscheidungen treffen, welche sich direkt auf den Menschen auswirken und eine große Verantwortung darstellen, steigt die Forderung, dass diese Entscheidungen nach gängigen ethischen Standards getroffen werden. Um dies zu gewährleisten, muss ermittelt werden, ob bestimmte Konzepte relevant für die Entscheidungen des neuronalen Netzes sind.
Die Arbeit [17] zeigt, wie für Menschen verständliche Konzepte, als lineare Kombination tiefer Neuronen in einem Vektor dargestellt werden. Diese Vektoren werden Konzept-Aktivierungs-Vektor/en (KAV) genannt und können mithilfe von Datenbeispielen ermittelt werden.
Die KAV ermöglichen es die Relevanz eines Konzeptes für eine bestimmte Entscheidung zu errechnen. Dies ermöglicht

es, Entscheidungen von neuronalen Netzen auf ethische Korrektheit [4] zu überprüfen.
In diesem Experiment wird vorerst nur das Erstellen von KAV und nicht das Errechnen dessen Relevanz betrachtet.

A. Formulierung der Hypothese

Mithilfe der Beschreibung eines Konzeptes, mittels eines Datensatzes, kann innerhalb des neuronalen Netzes ein Konzept-Aktivierungs-Vektor gefunden werden, welcher eine Repräsentation des Konzeptes darstellt. [5]

B. Experimentaufbau

Für dieses Experiment wird ein vortrainiertes Faltungsnetzwerk, mehrere Datensätze zur Beschreibung verschiedener Konzepte, sowie die dazugehörigen Referenzdatensätze benötigt. Um die ermittelten KAV visuell darzustellen, wird eine geeignete Visualisierungsmethode gewählt. Zudem ist ein menschlicher Experte vonnöten, um die Visualisierung zu interpretieren und damit das Konzept des KAV zu evaluieren. Da in diesem Experiment, anders als im Abschnitt III beschrieben, mit einer Interpretationsmethode anstelle eines Datensatzes evaluiert wird, sollte dessen Nützlichkeit bereits belegt wurden sein.

C. Experimentdurchführung

Zum Beschreiben eines Konzeptes mithilfe eines KAV werden zwei Datensätze P und N erstellt. Die Daten von P beschreiben das Konzept während N als Referenzdatensatz dient [6]. Durch das Anwenden des neuronalen Netzes auf beide Datensätze und das Speichern der Aktivierungen der verschiedenen Schichten [7], werden die Datensätze P_l und N_l erstellt, welche die Aktivierungsvektoren für die Schicht l enthalten.
Mit diesen Daten wird für jede Schicht des neuronalen Netzes ein linearer Klassifikator trainiert, welcher lernt das besagte Konzept zu erkennen. Die Gewichte des Klassifikators, der am besten klassifiziert, werden als KAV genommen (für eine genauere Beschreibung siehe [17]).
Anschließend wird der KAV mithilfe der Visualisierungsmethode visualisiert und vom menschlichen Experten auf dessen Richtigkeit überprüft.
Sofern diese Interpretationsmethoden ein gezieltes Lokalisieren von Konzepten ermöglicht, können Entscheidungen von neuronalen Netzen, mithilfe der KAV, auf ihre ethische Korrektheit überprüft werden.

[4] Mit ethischer Korrektheit ist hier gemeint, dass neuronale Netze keine Konzepte (z.b. Geschlecht, Herkunft o.a.) für ihre Entscheidungsfindung verwenden, welche aus ethischen Gründen, nicht ausschlaggebend für diese Entscheidungen sein dürfen.

[5] Diese Hypothese erfüllt zwar nicht die theoretischen Anforderungen an Interpretationsmethoden, doch mithilfe der KAV kann auf ethische Korrektheit getestet werden und eventuell wissenschaftliche Erkenntnisse gewonnen werden.

[6] Als Referenzdatensatz sollte N eine möglichst hohe Datenvarianz, nicht aber das besagte Konzept, abdecken.

[7] Die Aktivierungen von Faltungsschichten werden in einen eindimensionalen Vektor transformiert.

VII. Weiterführende Arbeiten

Um das Gelingen der verschiedenen Experimente zu erhöhen, ist es wichtig, geeignete Visualisierungs- und Attributionsmethoden zu verwenden. Hierzu gibt es bereits viele Forschungen und es sollten nur solche Methoden ausgewählt werden, bei denen noch keine Schwachstellen offengelegt wurden, welche die Methode für den praktischen Einsatz unbrauchbar machen. Bei den Feature Visualisierungsmethoden ist es zudem wichtig, dass diese, aussagekräftige Resultate liefern, auf dessen Basis Annahmen über das neuronale Netz oder den Trainingsdatensatz aufgestellt werden können. Um dies zu gewährleisten ist eine Evaluation und eventuell auch eine Optimierung einer geeigneten Methode vonnöten. Bei lernbasierten Feature Visualisierungsmethoden [4] können beispielsweise die Regularisierungsverfahren und viele Parameter optimiert werden, um die Aussagekraft ihrer Resultate zu verbessern.

Literatur

[1] Olah, Satyanarayan, Johnson, Carter, Schubert, Ye, Mordvintsev: The Building Blocks of Interpretability. In: Distill (2018)
[2] Carter, Armstrong, Schubert, Johnson, Olah: Exploring Neural Networks with Activation Atlases. In : In: Distill (2019)
[3] Erhan, Bengio, Courville, Vincent: Visualizing Higher-Layer Features of a Deep Network. In: Technical Report Univeristé de Montréal (2009)
[4] Olah, Mordvintsev, Schubert: Feature Visualization. In: Distill (2018)
[5] Mahendran, Vedaldi: Understanding Deep Image Representations by Inverting Them. (2014)
[6] Erhan, Bengio, Courville, Vincent: Visualizing Higher-Layer Features of a Deep Network. In : Technical Report, Univeristé de Montréal (2009)
[7] Simonyan, Vedaldi, Zisserman: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2014)
[8] Zeiler, Fergus: Visualizing and Understanding Convolutional Networks. (2013)
[9] Kindermans, Schutt, Alber, Muller, Erhan, Kim, Dahne: Learning how to explain neural networks: PatternNet and PatternAttribution. (2017)
[10] Fong, Vedaldi: Interpretable Explanations of Black Boxes by Meaningful Perturbation. (2018)
[11] Alsallakh, Jourabloo, Ye, Liu, Ren: Do Convolutional Neural Networks Learn Class Hierarchy? (2017)
[12] Kahng, Andrews, Kalro, Chau: ACTIVIS: Visual Exploration of Industry-Scale Deep Neural Network Models. In: IEEE Transactions on Visualization and Computer Graphics (2018)
[13] Doshi-Velez, Kim: Towards A Rigorous Science of Interpretable Machine Learning. (2017)
[14] Lipton: The Mythos of Model Interpretability. In CoRR (2017)
[15] Mordvintsev, Olah, Tyka: Inceptionism: Going Deeper into Neural Networks. (2015)
[16] Goodman, Flaxman: European Union Regulations on Algorithmic Decision Making and a "Right to Explanation" (2016)
[17] Anonymous authors: TCAV: Relative concept importance testing with linear concept activation vectors. Under review as a conference paper at ICLR (2018)