Stichprobenbasiertes Testen eines CNNs mithilfe von LRP

Eike Hannes Meyer, Email: *epsilonhmeyer@gmail.com*

28. November 2022

Motivation und Kontext. Künstliche neuronale Netze (NNs) und andere Methoden der künstlichen Intelligenz (KI) werden heutzutage in vielen Bereichen eingesetzt. Sie werden oft als Entscheidungshilfen in Anwendungsbereichen verwendet, die Menschen direkt oder indirekt betreffen, wie etwa bei der Auswertung von Bildmaterial für die Polizei oder in der Medizin [11]. Aufgrund dieses Einflusses ist es wünschenswert und verständlich, dass das Verhalten der verwendeten NNs genauer untersucht werden sollte, um ein gewisses Maß an Vertrauen in sie aufzubauen. Die Realität zeigt bereits, dass KI Systeme genauer untersucht werden müssen, um fehlerhaftes Verhalten aufzudecken. In den USA wird von Gerichten die Software *COMPAS* [6] verwendet, um das Risiko einzuschätzen, ob Angeklagte zu Wiederholungstätern werden. Diese auf KI basierte Software schreibt Schwarzen oft höhere Rückfallrisiken als anderen Ethnien zu, obwohl Beobachtungen in der Realität zeigen, dass dies nicht der Fall ist [4].
Im Gegensatz zu traditioneller Software, deren Verhalten explizit codiert ist und sich beispielsweise anhand des Kontrollflusses nachvollziehen lässt, ist das Verhalten von neuronalen Netzen implizit definiert und dadurch kaum durch Menschen nachvollziehbar. Dieses Verhalten ergibt sich aus den verwendeten Daten, die für die Berechnung der Gewichte von komplexen Graphen verwendet werden, deren Knotenanzahl teilweise in die Millionenbereiche geht. Aus diesem Grund können traditionelle Testmethoden für traditionelle Software kaum angewandt werden und neue Arten des Testens müssen gefunden und evaluiert werden. Mögliche Kandidaten hierfür entstammen dem Forschungsbereich der *explainable artificial intelligence* (XAI), der sich unter anderem mit der Nachvollziehbarkeit und Erklärung von getroffenen Entscheidungen neuronaler Netze beschäftigt [3].

Ziel dieser Arbeit ist es, die Methode *layer-wise relevance propagation* (LRP) aus dem Bereich der XAI zu untersuchen und mithilfe ausgewählter und gezielt veränderter Testdaten das Verhalten neuronaler Netze genauer zu prüfen. Im Detail wurde untersucht, ob sich mithilfe dieser Methoden unerwünschte oder falsche Korrelationen in den Trainingsdaten finden lassen, ohne eine aufwändige manuelle Durchsicht der großen Menge an Trainingsdaten durchführen zu müssen. Die Sicht auf diese Methoden erfolgt somit aus der Perspektive des Software Testings. Am Beispiel eines Netzes zur Bildklassifikation wird untersucht, inwieweit sich LRP als Blackbox Testmethode anwenden lässt und eignet.

Versuchsaufbau und Durchführung. Als Untersuchungsobjekt wurde ein *convolutional neural network* (CNN) zur Bildklassifikation gewählt. Bildklassifikation bezeichnet die Zuordnung eines Eingabedatums zu genau einer Klasse, deren Gesamtmenge vorher festgelegt wurde. Ein CNN ist eine spezielle Ausprägung eines neuronalen Netzes, das räumliche Nähe einzelner Werte innerhalb eines Eingabedatums mit berücksichtigt. Dieses Netz wurde auf einer Teilmenge des Datensatzes *ImageNet* trainiert, um so sein Verhalten zu erlernen. Dieser Datensatz wird vielfach als Benchmark für NNs verwendet [7]. *ImageNet* enthält 1000 verschiedene Klassen und ca. 1,2 Millionen verschiedene Bilder.
Das für die folgenden Versuche verwendete Netz verfügte über insgesamt 47.212.874 Parameter und 20 Schichten. Der Grund für die Wahl dieser Architektur ist die höhere Erkennungsgenauigkeit auf dem gewählten Datensatz, gegenüber einem regulären NN, auf denselben Trainingsdaten. Seit 2012 gelten CNNs als state-of-the-art Technologie im Bereich Bildklassifikation, da diese beispielsweise für Benchmarkwettbewerbe durchgängig besser abschnitten, als traditionelle NNs [7].
Um die Entscheidungsgrundlage des untersuchten CNNs zu visualisieren wird die Methode *layer-wise relevance propagation* verwendet. Diese Methode wurde von Bach et. al. [1] vorgeschlagen und entworfen. Die Methode verwendet die internen Parameter eines bereits trainierten Netzes, um die Relevanz einzelner Pixel des Eingabedatums auf die vom Netz getroffene Entscheidung zu berechnen und in einer Heatmap darzustellen.
Dies ist bei weitem nicht die einzige XAI Methode zur Darstellung von Heatmaps, wurde aber bereits erfolgreich in anderen Gebieten angewendet. Im Gegensatz zu anderen verbreiteten Methoden ist die

entstehende Heatmap klarer zu erkennen, da einzelne Pixel statt gesamter Regionen im Bild markiert sind [2].

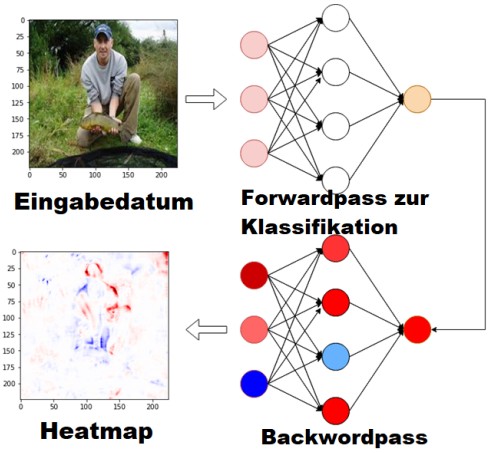

Abb. 1: Erstellung einer Heatmap mittels LRP

Die Auswahl der Testdaten erfolgte anhand mehrerer Kriterien. Als betrachtete Daten wurden Bilder der Klasse *tench* (Schleie) gewählt, da bei einer vorherigen Durchsicht festgestellt wurde, dass viele dieser Bilder große und stark ähnliche Artefakte enthalten. Mit Artefakten sind Bereiche und Objekte im Bild gemeint, die nicht direkt zur Klasseninstanz gehören. Eine weitere Rolle dabei spielten die Erkennungsgenauigkeit des Netzes. Hierfür wurden händisch diejenigen Bilder ausgesucht, betrachtet und verändert, für die das Netz jeweils die höchste und niedrigste Erkennungssicherheit angab, da diese Bilder scheinbar Eigenschaften aufweisen, die für die Klassifikation ein Rolle spielen.
Für die Versuche wurden sowohl Originalbilder aus *ImageNet* als auch händisch veränderte Bilder verwendet. Es wurden zwei Arten von manuellen Veränderungen durchgeführt:

- Die erste Art der Manipulation ist das Entfernen der eigentlichen Klasseninstanz aus dem Bild. Der verbleibende Bereich wurde mit Rauschen gefüllt. In diesem Fall die *Schleie* (Siehe Abb. 4)
- Die zweite Art der Manipulation ist das Entfernen von sämtlichen Bildteile, außer der eigentlichen Klasseninstanz. Wieder wurde der verbleibende Bereich mit Rauschen gefüllt. In diesem Fall alles außer der *Schleie* (Siehe Abb.6.)

Diese Manipulationen wurden vorgenommen, um genauer betrachten zu können, welche Teile der Bilder vom CNN verwendet werden, um eine Klassifikation vorzunehmen. Implizit soll also mit untersucht werden, welche Charakteristika das Netz erlernt hat, um

Abb. 2: ILSVRC2012_00009379.JPEG.aus [7]

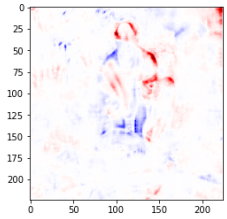

Abb. 3: Die mit LRP erzeugte Heatmap desselben Bildes

Klassen zu identifizieren. Der Grund für die Verwendung von Rauschen ist, dass in der Arbeit [10] gezeigt wurde, dass Rauschen in den Trainingsdaten einen positiven Einfluss auf die Abstraktionsfähigkeit und Gesamtperformance des Netzes haben kann und deshalb auch Testdaten auf eine potentiell positive Art verändert werden sollten, statt auf eine Art, die möglicherweise negative Einflüsse haben kann. Für alle Versuche wurden dieselben Bildinstanzen verwendet, um die Beobachtungen konsistent zu halten.

Beobachtungen und Diskussion. In diesem Abschnitt werden die Ergebnisse des vorher beschriebenen Versuches dargestellt und diskutiert [5]. Dargestellt werden die verwendeten Bilder und die daraus mithilfe vom verwendeten CNN und LRP erzeugte korrespondierende Heatmap. Rote Pixel in diesen Heatmaps zeigen Pixel an, die besonders relevant für die Klassifikation des Bildes durch das CNN waren. Blaue Pixel zeigen an, wenn Pixel gegen die gegebene Entscheidung sprachen. Weiße Pixel sind neutral und hatten keinen Einfluss auf die Entscheidung.
Im ersten Versuch wurden verschiedene, unveränderte Bilder aus *ImageNet*, die nicht Teil der Trainingsdaten waren, verwendet und eine Heatmap mittels LRP erzeugt.

Das unveränderte Bild 2 wurde mit einer Sicherheit von 45,53% korrekt als *tench* klassifiziert. Auf der dazugehörigen Heatmap ist zu sehen, wie der Kopf und die Knie des Menschen relevant für die Klassifikation waren. Weiterhin scheint ein Teil der Birken im Hintergrund relevant für die korrekte Klassifikation zu sein. Obwohl der Fisch die eigentliche Klasseninstanz ist, sprach der obere Teil sogar gegen eine Klassifikation als *tench*.

Abb. 4: ILSVRC2012_val_00037861.JPEG aus [7]

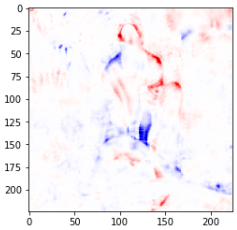

Abb. 5: Die durch LRP erzeugte Heatmap desselben Bildes

Das Bild 4 wurde mit einer Sicherheit von 51,34% korrekt klassifiziert. Im Vergleich zum Bild sieht man, dass das Ersetzen der eigentlichen Instanz dem Netz sogar half, das Bild korrekt zu klassifizieren. Der einzige Unterschied der relevanten Teile zum Bild ist, das die rechte Hand des Menschen nun einen positiven Beitrag zur korrekten Klassifikation leistet und dass das linke Bein deutlicher dazu beiträgt.

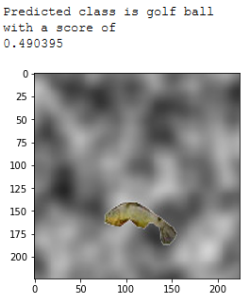

Abb. 6: ILSVRC2012_00009379.JPEG.aus [7]

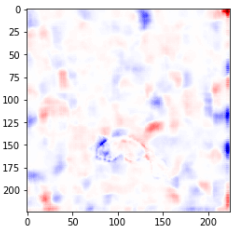

Abb. 7: Die mit LRP erzeugte Heatmap desselben Bildes

Das Bild 6 wurde mit einer Sicherheit von 49% falsch als *golf ball* klassifiziert. In der Heatmap ist der Umriss der eigentlichen Instanz noch sichtbar und spricht in etwa gleichen Teilen sowohl für als auch gegen diese Klassifizierung. Interessant ist, dass das Modell große Teile des verrauschten Hintergrundes zur Klassifikation stärker mitbenutzt, als in den bisher betrachteten Bildern.

Abgrenzung und Related work. Im Gegensatz zu theoretischen Arbeiten [1], aus denen die hier verwendete Methode stammt, erfolgt die Sicht als die eines Softwaretesters und aus Optimierungssicht. Diese Sicht erfolgt aus einer Blackbox Perspektive, von einigen Autoren vorgestellte Whitebox Methoden wurden nicht betrachtet [9].
In einer dieser ähnlichen Arbeiten [2] wurde auch LRP verwendet um das Fehlverhalten eines Netzes und Probleme mit den Trainingsdaten aufzudecken. Hier wurden jedoch nur unveränderte Bilder verwendet und die Ergebnisse nicht durch weitere Experimente untermauert.

Fazit und Ausblick. Mithilfe der Versuche in dieser Arbeit konnte eine Art Fehlverhalten des CNNs identifiziert werden. Durch die Verwendung von LRP auf ausgewählte und händisch manipulierte Datenbeispiele konnte ermittelt werden, dass das verwendete CNN Bilder der Klasse *tench* nicht an der Klasseninstanz selber, sondern an Artefakten im Bild, die nicht zur eigentlichen Klasse selber gehören, erkennt. Besonders deutlich wurde dies an den Stichproben, in denen die Klasseninstanz entweder komplett fehlte, oder aber nur diese Instanz vorhanden war. Im ersten Fall klassifizierte das CNN das Bild dennoch genau der Klasse zugehörig, aus der das Originalbild stammte. Im zweiten Fall wurde das Bild vollkommen falsch klassifiziert, wenngleich nur die Klasseninstanz im Bild vorhanden war.
Obwohl die Anwendung von LRP bereits mit wenigen gezielten Stichproben, wie beim Testen traditioneller Software, Probleme aufzeigen konnte, ist dieser Vorgang dennoch aufwändig. Jede Klasse, die das CNN erkennen soll, muss einzeln untersucht werden. Um den Aufwand zu minimieren und bessere Ergebnisse zu erhalten, sollte die Auswahl dieser Stichproben von Experten für die jeweilige Anwendungsdomäne erfolgen. Ebenso sollte vorher, wie bei traditioneller Softwareentwicklung, eine möglichst genaue Funktionsbeschreibung des NNs aufgestellt werden, sodass gegen diese Anforderungen getestet werden und die Auswahl des Testdaten spezifischer erfolgen kann.
Die Ergebnisse sind nur eingeschränkt verallgemeinerbar, einmal durch die Verwendung genau einer Art von KI und einem Datensatz, sowie nur der Aufgabe der Bildklassifikation. Während die Methode LRP auf beliebig große NNs und CNNs erfolgreich angewandt werden kann [1], sollten dennoch andere Arten von Netzen, Daten und Anwendungsfällen betrachtet werden. Ebenso sollten weitere Arten der Bildmanipulation verwendetet werden, wie beispielsweise die Arten von *data augmentation*, wie sie mittlerweile nahezu immer vor dem Training eines Netzes erfolgt [8]. Die Art der Testdatenauswahl sollte ebenso verfeinert werden, wie etwa durch die Erkennung von Anomalien in den verwendeten Daten.
Folgende weiter Handlungsempfehlungen ergeben sich aus den Versuchen: Um die Qualität eines Netzes zu beurteilen sollte nicht nur die Genauigkeit der Erkennung verwendet werden, da diese hoch sein kann, obwohl das Netz nicht die eigentliche Klasseninstanz zur Klassifikation verwendet worden sein kann. Dies kann genauer durch die in den Versuchen dargestellten Methoden erfolgen. Weiterhin kann es hilfreich sein, so veränderte Bilder auch für das Training zu verwenden. In der Arbeit [10] wurde nachgewiesen, dass Rauschen in den Trainingsdaten einen positiven Einfluss auf die Abstraktionsfähigkeit und Gesamtperformance des Netzes haben können. Hierbei wäre eine Automatisierung der Augmentations sinnvoll und hilfreich.

Literatur

[1] S. Bach, Alexander Binder, Grégoire Montavon, F. Klauschen, K. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015.

[2] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *CoRR*, abs/1902.10178, 2019.

[3] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016.

[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[5] Eike Hannes Meyer. Evaluation von visualisierungsmethoden als testverfahren für cnns am beispiel bildklassifikation. https://users.informatik.haw-hamburg.de/ ubicomp/arbeiten/bachelor/meyer.pdf, 2020.

[6] Northpointe. Compas. https://psrac.bja.ojp.gov.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[8] Connor et al.: Shorten. A survey on image data augmentation for deep learning. Journal of Big Data,6–60 (2019).

[9] Youcheng Sun, Xiaowei Huang, Daniel Kroening, James Sharp, Matthew Hill, and Rob Ashmore. Testing deep neural networks, 2019.

[10] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization, 2018.

[11] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated, 2020.