Analyse von Saliency Map Methoden mittels Adversarial Examples

Nicole Wagner

Hochschule Darmstadt - Fachbereiche Mathematik und Naturwissenschaften & Informatik

Einleitung

Für die Bildverarbeitung und -klassifikation sind Convolutional Neural Networks (CNNs) zu einer Schlüsseltechnologie geworden, die beeindruckende Erfolge in verschiedensten Anwendungen erzielt hat. Saliency Maps sind ein bekannter Ansatz, um Einblicke in die Entscheidungsprozesse von CNNs zu gewinnen, indem sie darstellen, welche Pixel eines Bildes für die Vorhersage eines Modells am relevantesten sind.

Um die Qualität zu messen, werden Saliency Maps von Bildern und deren Adversarial Examples erzeugt und miteinander verglichen. Erzeugt ein Adversarial Example eine ähnliche Saliency Map wie sein Originalbild, so hat das Erklärmodell nicht auf das veränderte Bild und dessen Klassifikation reagiert, was auf eine geringe Aussagekraft hinweist. Zur Messung der Ähnlichkeit von Saliency Maps zwischen den Originalbildern und den Adversarial Examples wird die Wasserstein-Metrik als Kennzahl eingeführt. Sie misst die Ähnlichkeit zweier Verteilungen. Die Saliency Maps werden dafür entsprechend skaliert. Darüber hinaus wird das bekannte Relevance Ranking und der Spearman-Rangkorrelationskoeffizient verwendet.

Für die Analyse wird ein, auf den CIFAR10-Datensatz trainiertes, CNN verwendet, das zusammen mit den zugehörigen Adversarial Examples von [2] bereitgestellt wurde.

Adversarial Examples

Für Adversarial Examples werden Bilder, die vom CNN klassifiziert werden können, minimal verändert, um ein anderes Klassifikationsergebnis zu erhalten. Für Menschen können die Änderungen zwar sichtbar sein, doch die Klasse des Bildes ist weiterhin eindeutig zu erkennen. Adversarial Examples werden durch ein lineares Optimierungsproblem ermittelt. Aufgrund der hohen Komplexität der Berechnung werden sie durch Approximationsverfahren berechnet. Die hier untersuchten Adversarial Examples von [2] wurden mittels L-BFGS Attack erstellt.

Saliency Map Methoden

Saliency Maps zeigen die Relevanz einzelner Pixel oder Features eines Bildes für die Klassifikation in CNNs anhand von Relevance Scores.

Vanilla Gradient

Vanilla Gradient berechnet die Gradienten der Score-Funktion S_c für ein gegebenes Bild I_0 und die gewünschte Klasse c [5]. Jedem Pixel wird der größte absolute Gradient der Farbwerte als Relevance Score zugeordnet. Der Gradient soll ein Indikator für die Relevanz eines Pixels sein, da eine kleine Änderung von Pixeln mit einem hohen absoluten Gradienten einen großen Einfluss auf den Klassen-Score hat.

Grad-CAM

Ziel des Grad-CAM Verfahrens ist es, diejenigen Regionen im Bild zu identifizieren, die im letzten Layer des CNN aktiviert werden. Dafür werden die Feature Maps des letzten Convolution Layers und deren Gradienten bezüglich der Score-Funktion betrachtet [4].

Layer-wise Relevance Propagation

Bei der Layer-wise Relevance Propagation (LRP) werden die Relevance Scores jedes Pixels für eine Klasse c mit einem sogenannten Backwardpass berechnet. Dafür wird das Ergebnis der Score-Funktion rückwärts durch das CNN geschickt und immer weiter zerlegt, bis jeder Pixel einen eigenen Relevance Score erhält [1]. Es wurden die Zerlegungsregeln $\alpha\beta$-Regel und ε-Regel mit den Parametern $\alpha = 1$, $\alpha = 2$ und $\varepsilon = 1$ untersucht.

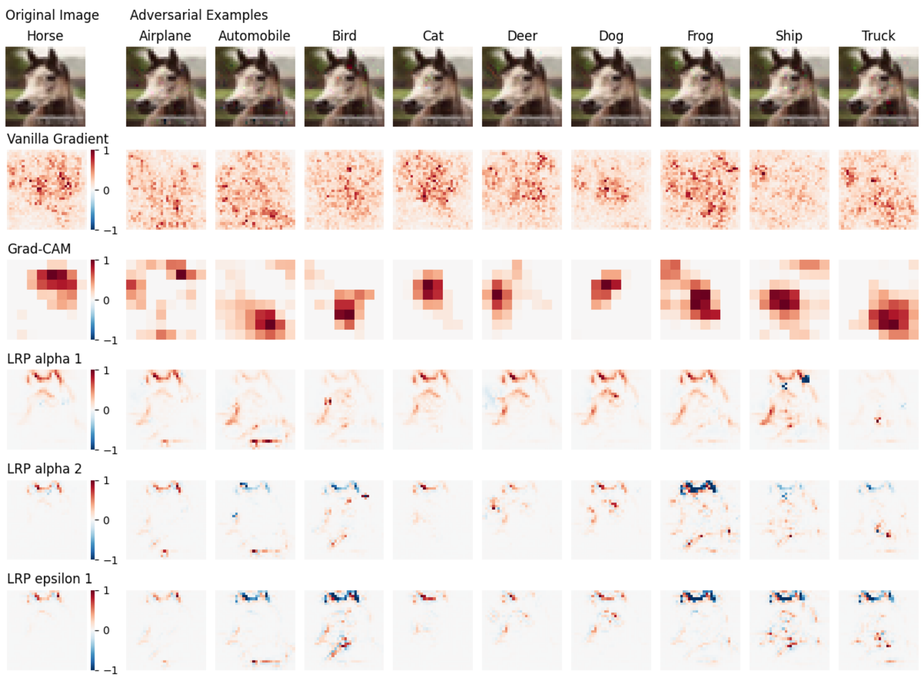


Abbildung 1. Visuelle Darstellung der Saliency Maps für alle Verfahren anhand des Beispielbildes Pferd

Statistische Methoden und Kennzahlen

Die statistischen Kennzahlen Vergleichen die Saliency Maps der Originalbilder mit den Saliency Maps der zugehörigen Adversarial Examples und quantifizieren deren Ähnlichkeit.

Spearman-Rangkorrelationskoeffizient

Der Spearman-Rangkorrelationskoeffizient ist ein Maß für den monotonen Zusammenhang zweier Stichproben. In diesem Fall handelt es sich bei den Stichproben um die Relevance Scores der Pixel. Der Spearman-Rangkorrelationskoeffizienten wird dabei für jedes Originalbild und dessen Adversarial Examples berechnet. Ein großer positiver/negativer Wert deutet auf einen positiven/negativen Zusammenhang hin.

Relevance Ranking

Die Relevanzverschiebung für ein Pixel i ergibt sich aus der Differenz der Ränge $d'_i = s'_i - r'_i$. Negative Werte bedeuten somit eine höhere Relevanz des Pixels im Adversarial Example. Positive Werte bedeuten eine niedrigere Relevanz des Pixels im Adversarial Example. Für die statistische Auswertung wird die absolute Relevanzverschiebung des Pixels mit dem größten Relevance Score im Originalbild $x_{(1)}$ zu jedem Adversarial Example bestimmt. Ein kleiner Wert deutet darauf hin, dass der Pixel im Adversarial Example auch eine hohe Relevanz hat.

Statistische Methoden und Kennzahlen

Wasserstein-Metrik

Die Wasserstein-Metrik misst die Ähnlichkeit zwischen zwei Verteilungen. Hierbei wird der minimale Aufwand berechnet, der erforderlich ist, um die Masse von einer Verteilung in die Masse der anderen Verteilung zu transportieren. Die Metrik berechnet sich aus der Lösung des Transportproblems [3]. Ist der Wert klein, so sind die Verteilungen ähnlich.

Definition 1 (Transportproblem)**.** Seien $\mathcal{I} = \{x_i | i \in I\}$ die Menge an Quellen und $\mathcal{J} = \{y_j | j \in J\}$ die Menge der Ziele. Die Masse einer Quelle sei v_i und die Masse eines Ziels sei w_i. Seien d_{ij} die Kosten für den Transport von Masse von x_i nach y_j. Dann wird das Transportproblem definiert durch

$$
\begin{aligned}
\min \quad & \sum_{i \in I} \sum_{j \in J} d_{ij} f_{ij} \\
\text{s.t.} \quad & f_{ij} \geq 0, && \forall i \in I, j \in J \\
& \sum_{j \in J} f_{ij} = v_i, && \forall i \in I \\
& \sum_{i \in I} f_{ij} = w_j, && \forall j \in J
\end{aligned}
$$

Abbildung 2. Darstellung von zwei diskreten zweidimensionalen Verteilungen

wobei f_{ij} den Fluss von Masse der Quelle x_i an das Ziel y_j beschreibt.

Definition 2 (Wasserstein-Metrik)**.** Gegeben sei die Lösung des Transportproblems aus Definition 1. Dann berechnet sich die Wasserstein-Metrik zwischen zwei Verteilungen von Masse v und w durch

$$\mathrm{WS}(v, w) = \frac{\sum_{i \in I} \sum_{j \in J} d_{ij} f_{ij}}{\sum_{i \in I} \sum_{j \in J} f_{ij}} = \frac{\sum_{i \in I} \sum_{j \in J} d_{ij} f_{ij}}{\sum_{j \in J} w_j}.$$

Für den Vergleich der Saliency Maps werden die Relevance Scores in eine Verteilung umgewandelt. Für die LRP Saliency Maps ist es möglich, dass negative Relevance Scores auftreten. Sie werden für die Auswertung auf Null gesetzt. Die positiven Relevance Scores werden skaliert, sodass ihre Summe 1 ergibt. Die Wasserstein-Metrik wird bezüglich zweier Distanzmaße berechnet.

1. **Euklidische Distanzen**: Für jeden Relevance Score sei die Position auf der Ebene anhand der Koordinaten der Pixel definiert. Als Distanzmaß zwischen zwei Pixeln wird die euklidische Distanz der Koordinaten gewählt $d_{ij} = \|x_i - y_j\|_2$.
2. **Rangdifferenzen als Distanz**: Es werden getrennt für die skalierten Relevance Scores des Originalbildes $v'_1, \ldots, v'_n$ und die skalierten Relevance Scores des Adversarial Examples $w'_1, \ldots, w'_n$ die Ränge gebildet. Die Differenz der Ränge bildet die Distanz zwischen den Pixeln.

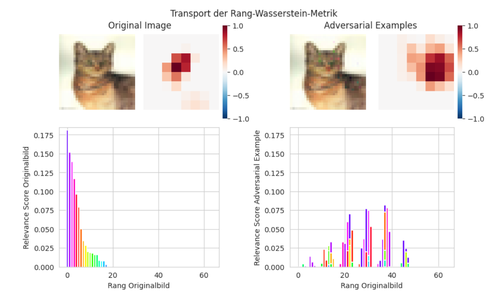


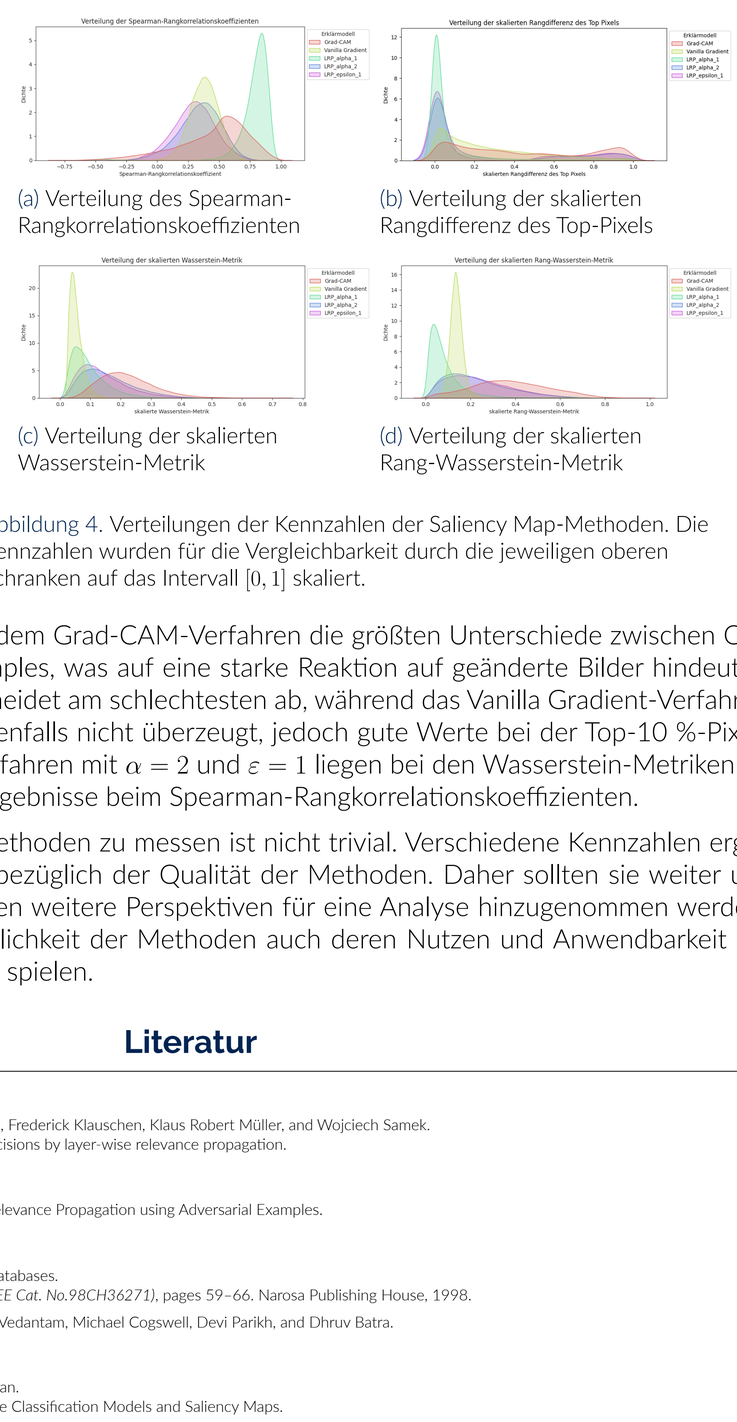
Abbildung 3. Darstellung des Transports bei der Rang-Wasserstein-Metrik. Die x-Achse zeigt in beiden Fällen die Ränge des Originalbildes. Die y-Achse zeigt die Relevance Scores. Die Farben zeigen die Quellpixel im Originalbild.

Ergebnisse

Die verschiedenen Methoden erzeugen sehr unterschiedliche Saliency Maps. Sie unterscheiden sich in der Größe der verteilten Relevance Scores und der Verteilung der Relevance Scores auf die Pixel. Während das Vanilla Gradient-Verfahren die Relevanz auf viele Pixel verteilt, konzentrieren sich die LRP-Verfahren auf wenige relevante Pixel. Das Grad-CAM-Verfahren erzeugt Saliency Maps der Größe 8×8 und die Anzahl der relevanten Regionen in den Saliency Maps variiert am meisten.

Für die Analyse der Saliency Maps mit Adversarial Examples wurden die oben stehenden statistischen Kennzahlen für alle Paare von Adversarial Example und Originalbild berechnet. Abbildung 4 zeigt je Kennzahl für jede Saliency Map-Methode die Verteilung dieser Ergebnisse. Für die Vergleichbarkeit zwischen den Methoden wurden die Wasserstein-Metriken geeignet skaliert.

Der Vergleich der Saliency Map-Methoden anhand verschiedener Metriken führt zu unterschiedlichen Ergebnissen.



(a) Verteilung des Spearman-Rangkorrelationskoeffizienten

(b) Verteilung der skalierten Rangdifferenz des Top-Pixels

(c) Verteilung der skalierten Wasserstein-Metrik

(d) Verteilung der skalierten Rang-Wasserstein-Metrik

Abbildung 4. Verteilungen der Kennzahlen der Saliency Map-Methoden. Die Kennzahlen wurden für die Vergleichbarkeit durch die jeweiligen oberen Schranken auf das Intervall $[0, 1]$ skaliert.

Die Wasserstein-Metrik zeigt bei dem Grad-CAM-Verfahren die größten Unterschiede zwischen Originalbildern und Adversarial Examples, was auf eine starke Reaktion auf geänderte Bilder hindeutet. Das LRP-Verfahren mit $\alpha = 1$ schneidet am schlechtesten ab, während das Vanilla Gradient-Verfahren bei den Wasserstein-Metriken ebenfalls nicht überzeugt, jedoch gute Werte bei der Top-10 %-Pixel-Schnittmenge erzielt. Die LRP-Verfahren mit $\alpha = 2$ und $\varepsilon = 1$ liegen bei den Wasserstein-Metriken im Mittelfeld, erzielen jedoch gute Ergebnisse beim Spearman-Rangkorrelationskoeffizienten.

Die Qualität von Saliency Map-Methoden zu messen ist nicht trivial. Verschiedene Kennzahlen ergeben unterschiedliche Ergebnisse bezüglich der Qualität der Methoden. Daher sollten sie weiter untersucht werden. Zusätzlich können weitere Perspektiven für eine Analyse hinzugenommen werden. Dabei könnten neben der Verlässlichkeit der Methoden auch deren Nutzen und Anwendbarkeit für den Menschen eine zentrale Rolle spielen.

Literatur

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek.
On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.
PLoS ONE, 10(7), 7 2015.

[2] Tamara R. Dieter and Horst Zisgen.
Evaluation of the Explanatory Power Of Layer-wise Relevance Propagation using Adversarial Examples.
Neural Processing Letters, 55(7):8531–8550, 12 2023.

[3] Y. Rubner, C. Tomasi, and L.J. Guibas.
A metric for distributions with applications to image databases.
In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66. Narosa Publishing House, 1998.

[4] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra.
Grad-CAM: Why did you say that?
ArXiv, abs/1611.07450, 11 2016.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman.
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.
CoRR, 12 2013.