Explainable AI

Christoph Würsch, ICE

15.4.2024

1 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) ist eine Technik zur Visualisierung der Entscheidungen von Convolutional Neural Networks (CNNs). Sie wird verwendet, um die Bereiche eines Eingabebildes zu identifizieren, die wesentlich zur Klassifikation beigetragen haben.

Ein CNN führt zuerst einen Vorwärtsdurchlauf mit dem Eingabebild durch, was in einer Reihe von Feature-Maps am Ausgang einer spezifischen Convolution-Schicht resultiert. Diese Maps, bezeichnet als A^k, repräsentieren verschiedene Aspekte des Bildes.

Nach der Klassenvorhersage erfolgt ein Rückwärtsdurchlauf zur Berechnung der Gradienten der Ausgabe des Netzes in Bezug auf die Feature-Maps. Mathematisch wird dies als $\frac{\partial y^c}{\partial A^k}$ dargestellt, wobei y^c die Netzwerkausgabe für die Klasse c ist.

- **Berechnung der Gewichte** α_k^c: Die Gradienten werden über die räumlichen Dimensionen der Feature-Maps gemittelt, um ein Bedeutungsgewicht α_k^c für jede Map zu bestimmen:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

  Hier ist Z die Gesamtanzahl der Elemente in der Feature-Map.

- **Berechnung der Klasse-Aktivierungskarte**: Die finale Aktivierungskarte für die Klasse c wird durch eine gewichtete Summe der Feature-Maps, gewichtet mit ihren jeweiligen Bedeutungsgewichten, berechnet:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

  Die Anwendung der ReLU-Funktion stellt sicher, dass nur positive Beiträge berücksichtigt werden.

2 Gradienten-basierte XAI-Methoden

- **Layer-wise Relevance Propagation (LRP)** Layer-wise Relevance Propagation (LRP) zielt darauf ab, die Frage zu beantworten: "Welche Merkmale eines Eingabedatensatzes tragen wie stark zu einer bestimmten Vorhersage eines neuronalen Netzwerks bei?" LRP arbeitet, indem es Relevanzwerte, die die Vorhersage der Ausgabeschicht darstellen, rückwärts durch die Schichten des Netzwerks bis zum Eingabelevel überträgt. Diese Technik ermöglicht es, spezifisch zu verstehen, wie einzelne Input-Features das Ergebnis beeinflussen, was besonders in tiefen Netzwerken nützlich ist, wo die direkte Interpretation der Aktivierungen schwierig sein kann.

- **Deep Learning Important FeaTures (DeepLIFT)** DeepLIFT vergleicht die Aktivierung von Neuronen mit einem Referenzinput, um die Beiträge einzelner Neuronen zu bewerten. Diese Methode ermöglicht es, die Effekte von Input-Features auf die Ausgabe unabhängig von anderen Features zu identifizieren. Es ist besonders wertvoll, weil es dazu beiträgt, nicht nur die Richtung (positiv oder negativ) sondern auch das Ausmaß des Einflusses zu identifizieren. Diese Methode kann sogar in Fällen angewendet werden, in denen der Input stark miteinander korrelierte Features enthält, indem sie den jeweiligen Beitrag jedes Features unabhängig unterscheidet.

- **Saliency Maps** Saliency Maps nutzen die Gradienten der Ausgabe in Bezug auf das Eingabebild, um die Wichtigkeit jedes Pixels zu bestimmen. Diese Technik ist besonders nützlich, um visuell darzustellen, welche Teile eines Bildes die Vorhersagen eines Modells beeinflussen. Höhere Gradientenwerte bedeuten höhere Wichtigkeit. Sie sind einfach zu implementieren und bieten eine schnelle Methode zur Inspektion der Modellfunktionalität, sind jedoch oft durch Rauschen und die Empfindlichkeit gegenüber kleinen Veränderungen in der Eingabe beeinträchtigt.

- **Deconvolutional Networks (DeconvNets)** Deconvolutional Networks, auch DeconvNets genannt, sind eine Methode zur Visualisierung und Interpretation der Funktionen, die in den höheren Ebenen eines CNN aktiviert werden. Durch die Anwendung transponierter Convolution-Operationen, die als Umkehrung der normalen Convolution betrachtet werden können, ermöglichen sie die Rekonstruktion des Inputs, der zur Aktivierung spezifischer Features in der Netzwerkarchitektur führt. Obwohl nützlich, können DeconvNets manchmal irreführende Visualisierungen erzeugen, wenn die Rückprojektion zu stark von der tatsächlichen Feature-Aktivierung abweicht.

- **SmoothGrad** SmoothGrad verbessert Saliency Maps, indem es die Auswirkungen von Rauschen reduziert. Dies wird erreicht, indem der Durchschnitt vieler Saliency Maps berechnet wird, die aus leicht veränderten Versionen des ursprünglichen Inputs erzeugt wurden. Durch das Hinzufügen von zufälligem Rauschen zu mehreren Kopien eines Eingabebildes und das anschließende Durchschnittsbilden der resultierenden Saliency Maps wird das inhärente Rauschen, das oft in einzelnen Maps vorhanden ist, reduziert. Diese Methode macht die Interpretation der visuellen Daten zuverlässiger, erhöht jedoch die Rechenlast.

- **Guided Backpropagation (GBP)** GBP kombiniert Techniken der Backpropagation und Deconvolution, indem es negative Gradienten während des Rückwärtsdurchlaufs unterdrückt, um klarere Visualisierungen der Features zu erzeugen, die zur Entscheidungsfindung beitragen. Obwohl GBP ein nützliches Werkzeug für die Visualisierung und Verständnis der Entscheidungspfade in CNNs ist, kann es durch das Unterdrücken negativer Beiträge auch zu einem unvollständigen Bild der Netzwerkdynamik führen.

- **Integrated Gradients (IG)** Integrated Gradients bietet eine Methode zur Berechnung der Feature-Wichtigkeit durch die Integration der Gradienten entlang eines Pfades von einem Basiszustand zum tatsächlichen Eingabestatus. Diese Technik ist robust gegenüber vielen der Probleme, die in einfacheren Gradientenmethoden auftreten, wie Saturation oder die Unempfindlichkeit gegenüber bestimmten Inputs. Sie erfordert jedoch eine sorgfältige Auswahl des Basiszustandes und ist rechenintensiv.

- **Gradient-weighted Class Activation Mapping (Grad-CAM)** Grad-CAM nutzt die Gradienteninformationen von Zielklassen, die auf die letzten konvolutionellen Feature-Maps angewendet werden, um zu zeigen, welche Regionen eines Bildes am meisten zur Klassifizierung beitragen. Diese Methode bietet eine hervorragende Balance zwischen Hochleistungsbildlokalisierung und einfacher Implementierung, kann jedoch aufgrund ihrer Abhängigkeit von der letzten konvolutionellen Schicht in ihrer Fähigkeit begrenzt sein, feinere Details zu erfassen.

3 Vergleich von Gradienten-basierten XAI-Methoden

Methode	Beschreibung	Stärken und Schwächen
LRP	Zuweist relevante Scores entlang des Netzwerks rückwärts, basierend auf der Beitrag der einzelnen Neuronen zur Endvorhersage.	**Stärken:** Detailliertes Verständnis auf Neuronen-Ebene. **Schwächen:** Kann bei komplexen Netzwerken schwierig zu interpretieren sein.
DeepLIFT	Vergleicht die Aktivierung eines Neurons zu einem "Referenzinput" und berechnet Beiträge basierend auf den Unterschieden.	**Stärken:** Funktioniert gut bei Nichtlinearitäten. **Schwächen:** Auswahl eines geeigneten Referenzinputs kann herausfordernd sein.
Saliency	Nutzt die Gradienten der Ausgabe in Bezug auf das Eingabebild, um die Wichtigkeit jedes Pixels zu bestimmen.	**Stärken:** Einfach zu implementieren. **Schwächen:** Kann rauschanfällig sein; oft schwer zu interpretieren.

Methode	Beschreibung	Stärken und Schwächen
DeconvNets	Verwendet transponierte Convolution-Operationen, um die Feature-Aktivierungen im Netzwerk rückzuverfolgen.	**Stärken:** Visualisiert Feature-Aktivierungen. **Schwächen:** Kann irreführende Visualisierungen erzeugen.
SmoothGrad	Mittelt die Saliency Maps über viele leicht gestörte Versionen des Eingabebildes, um Rauschen zu reduzieren.	**Stärken:** Reduziert Rauschen in der Visualisierung. **Schwächen:** Rechenintensiv.
GBP	Kombiniert Backpropagation und Deconvolution, wobei negative Gradienten während des Rückwärtsdurchlaufs unterdrückt werden.	**Stärken:** Klare Visualisierungen. **Schwächen:** Kann wichtige negative Beiträge ausblenden.
IG	Integriert den Gradienten entlang eines Pfads vom Referenzzustand zum tatsächlichen Eingabestatus.	**Stärken:** Bietet detaillierte Attributionswerte, reduziert Saturationseffekte. **Schwächen:** Rechenintensiv und abhängig von der Wahl des Referenzpunktes.
Grad-CAM	Verwendet die Gradienten von Zielklassen, die auf die Feature-Maps einer CNN-Schicht angewendet werden, um eine Heatmap zu erzeugen.	**Stärken:** Gute Lokalisierung; einfach zu verstehen und zu implementieren. **Schwächen:** Begrenzt auf die visuellen Fähigkeiten der gewählten Convolution-Schicht.