Zusammenfassung

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.

Diese Masterarbeit untersucht die Qualität der drei Saliency Map-Methoden Vanilla Gradient, Grad-CAM und Layer-wise Relevance Propagation (LRP) mit den beiden Zerlegungsregeln $\alpha\beta$-Regel und ε-Regel mit den Parametern $\alpha = 1$, $\alpha = 2$ und $\varepsilon = 1$.

Um die Qualität zu messen, werden Saliency Maps von Bildern und deren Adversarial Examples erzeugt und miteinander verglichen. Adversarial Examples sind minimal veränderte Versionen von Bildern, die ein Modell zu fehlerhaften Vorhersagen verleiten. 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 wurde in dieser Arbeit die Wasserstein-Metrik als neue Kennzahl eingeführt. Darüberhinaus wurde das bekannte Relevance Ranking und der Spearman-Rangkorrelationskoeffizient verwendet.

Bei den Wasserstein-Metriken hat sich das Grad-CAM-Verfahren als beste Methode herausgestellt. Der Spearman-Rangkorrelationskoeffizient hingegen bewertet das LRP-Verfahren mit $\varepsilon = 1$ am besten. Das LRP-Verfahren mit $\alpha = 1$ schneidet beim Vergleich am schlechtesten ab. Ein weiteres Ergebnis dieser Masterarbeit ist, dass verschiedene Metriken zur Messung der Ähnlichkeit zwischen Original- und Adversarial Saliency Maps die Qualität der Saliency Map-Methoden unterschiedlich bewertet.

Abstract

For image processing and classification, Convolutional Neural Networks (CNNs) have become a key technology which has achieved impressive success in various applications. Saliency maps are a well-known approach to gain insights into the decision-making processes of CNNs by visualising which pixels of an image are most relevant for the prediction of the model.

This master's thesis analyses the quality of the three saliency map methods Vanilla Gradient, Grad-CAM and Layer-wise Relevance Propagation (LRP) with the two decomposition rules $\alpha\beta$-rule and ε-rule with the parameters $\alpha = 1$, $\alpha = 2$ and $\varepsilon = 1$.

To measure the quality, saliency maps of images and their adversarial examples are generated and compared with each other. Adversarial examples are minimally modified versions of images that lead a model to make incorrect predictions. If an adversarial example generates a similar saliency map to its original image, the explanatory model has not reacted to the altered image and its classification, which indicates a low informative value. To measure the similarity of saliency maps between the original images and the adversarial examples, the Wasserstein metric was introduced as a new key figure in this work. In addition, the well-known relevance ranking and the Spearman rank correlation coefficient were used.

For the Wasserstein metrics, the Grad-CAM method proved to be the best method. The Spearman rank correlation coefficient, on the other hand, rates the LRP method with $\varepsilon = 1$ as the best. The LRP method with $\alpha = 1$ performs worst in the comparison. Another result of this master's thesis is that different metrics for measuring the similarity between original and adversarial saliency maps evaluate the quality of the saliency map methods differently.