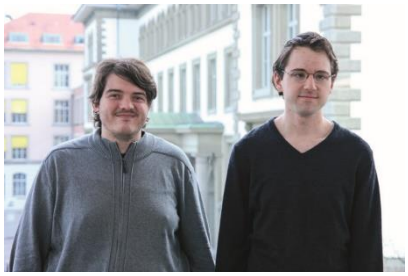
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Understanding Deep Neural Networks

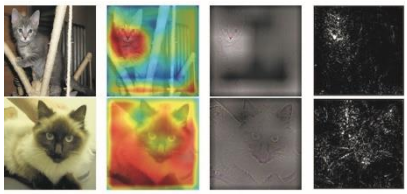


Diplomierende
David Kempf
Lino von Burg

Dozierende
Olaf Stern
Thilo Stadelmann

Künstliche Intelligenz (KI) umfasst eine Vielzahl von verschiedenen Problemlösungsmethoden. Während der letzten Dekade konnten in diesem Forschungsfeld sehr grosse Fortschritte verzeichnet werden und in der Folge hat KI sehr gute Ergebnisse in bisher Menschen vorbehaltenen Aufgabenbereichen erreicht. Insbesondere künstliche neuronale Netze (KNNs) gewannen im Zeitraum zwischen 2009 und 2012 zahlreiche internationale Wettbewerbe in maschinellem Lernen und erreichten sogar menschenähnliche Leistungen in Mustererkennungs-Anwendungen. Trotz dieser beeindruckenden Erfolge werden KNNs aufgrund der Komplexität der zugrundeliegenden Modelle oft als undurchsichtige 'Blackboxes' gesehen. Je mehr KNNs Teil der alltäglichen Prozesse werden, umso wichtiger wird es, ihre innere Funktionsweise zu verstehen. Diese Bachelorarbeit begutachtet deshalb den State of the Art von Debugging-Methoden für Convolutional Neural Networks (CNNs) und untersucht, wie hilfreich die Erklärungen von solchen Methoden für ein besseres Verständnis von Klassifikatoren sind.

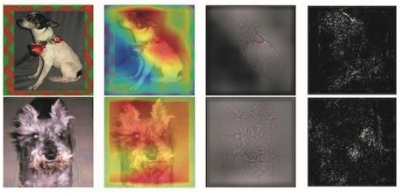
Im Rahmen dieser Arbeit wurde der State of the Art von Debugging-Methoden für CNNs untersucht und in eine neue Taxonomie eingeordnet. Die fundamentalen Ausprägungen dieser Methoden wurden auf ein CNN angewendet, das auf dem 'dogs vs cats' Wettbewerbsdatensatz trainiert worden ist.



Zwei Katzenbilder aus dem Datenset 'Dogs vs. Cats' von Kaggle. Die erste Zeile zeigt ein Beispiel einer nachvollziehbaren Erklärung, die zweite Reihe eine nicht nachvollziehbare. Es werden drei verschiedene Visualisierungen präsentiert.

Im Besonderen wurde zusätzlich zu dem Literatursurvey eine experimentelle Testumgebung für das Anwenden von Debugging-Methoden auf CNNs aufgebaut. Anhand dieser wurde ein Experiment für die Evaluation der begutachteten Methoden durchgeführt. Das Bilderkennungsproblem zur Unterscheidung von Katzen und Hunden ist Basis für die Untersuchung. Deren Ziel ist es, zu analysieren, wie das trainierte Modell ein Urteil fällt, wenn als Input für das Netzwerk Bilder eingegeben werden, die zu keiner der Outputklassen gehören.

Werden Debugging-Methoden benutzt, um die Funktion von CNNs besser zu verstehen, dann ist es wichtig, dass die Ergebnisse von diesen Methoden nicht nur erklären, wieso ein gegebenes Bild zu einer spezifischen Outputklasse zugeordnet wurde, sondern auch, wieso das gleiche Bild nicht einer anderen Outputklasse zugeordnet wurde. So fördern die Resultate das Verständnis, was für eine Vorstellung der Klassen das CNN aus den Trainingsdaten entwickelt hat. Daher wurde in einem Teil des Experiments verglichen, wie sich das Netzwerk bei Eingaben verhält, die nicht den Outputklassen entsprechen.



Zwei Hundebilder aus dem Datenset 'Dogs vs. Cats' von Kaggle. Die erste Zeile zeigt ein Beispiel einer nachvollziehbaren Erklärung, die zweite Reihe eines einer nicht nachvollziehbaren. Es werden drei verschiedene Visualisierungen präsentiert.