Explainable AI - Stand der Forschung und Technik

Studiengang: MAS | Vertiefung: MAS Data Science

Machine Learning ist zu einem regelrechten Hypethema geworden. Die Qualität der Ergebnisse ist für viele Anwendungsgebiete ausreichend geworden, so dass Menschen immer öfters mit den Entscheidungen von Algorithmen konfrontiert werden. Aber wie transparent arbeiten diese eigentlich? Weshalb wird mir ein bestimmter Film empfohlen oder, gravierender, ein Kredit verweigert? Explainable AI versucht bisher unverständliche Systeme für Menschen verständlich zu machen.

Ausgangslage

Die immer besseren Machine Learning (ML) Algorithmen in Kombination mit oftmals gigantischen Datenmengen ergeben Resultate, die für den produktiven Einsatz genügen und teilweise sogar die Leistungen von Menschen übertreffen. Aus dieser Sicht funktioniert ML einwandfrei. Problematisch wird deren Einsatz jedoch, wenn eine Entscheidung begründet werden soll. Der Verweis auf ein „durch ein computergestütztes Entscheidungssystem“ ist nicht befriedigend und verhindert manchmal den Einsatz von ML, beispielsweise in den Gebieten Medizin oder Justiz. Explainable AI (XAI) ist ein neues Forschungsgebiet, in welchem Methoden und Werkzeuge erarbeitet werden um das Verstehen und Begründen von ML basierten Entscheidungen zu ermöglichen.

Ziele

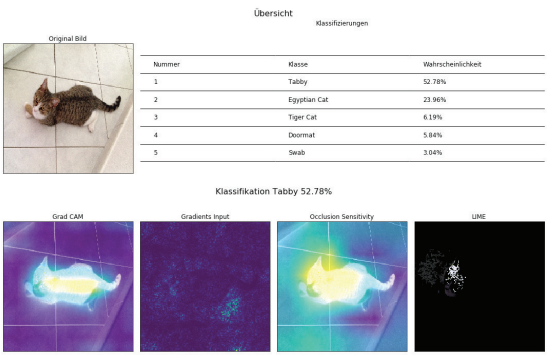
Diese Arbeit soll einen Überblick über vorhandene Methoden und Werkzeuge bieten, welche ML Modelle erklären und analysieren. In einigen Beispielen werden die Anwendungen dieser Techniken erläutert. Ein weiterer Aspekt sind die Rahmenbedingungen für zukünftige ML Anwendungen, die von XAI profitieren oder diese sogar voraussetzen.

Ergebnis

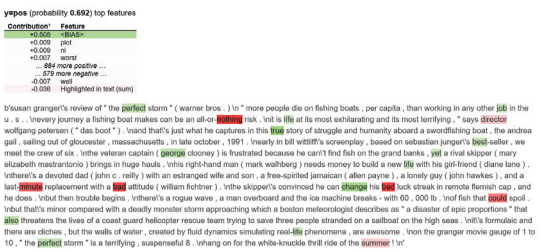
Durch die starke Nachfrage nach XAI, zum Teil gefordert durch Vorstösse von Politikern und Regierungsbehörden, ist die Anzahl neu entwickelter Methoden gross. Diese Arbeit hat einige bereits länger vorhandene Methoden an den Fallbeispielen Bilderkennung und Textanalyse angewendet. Dadurch konnte gezeigt werden, dass XAI das Verständnis für die Funktionsweise eines neuronalen Netzes und eines Random Forest Klassifikators erhöht. Durch ein absichtlich mit ungünstigen Daten erzeugtes Modell konnte auch das Prinzip des „Kluger Hans“-Effekt gezeigt werden und wie man solche Probleme erkennen und vermeiden kann.

Eine weitere Erkenntnis meiner Arbeit ist auch die Schwierigkeit eines Einsatzes dieser Werkzeuge im alltäglichen Betrieb. Zum einen sind die erzeugten Erklärungen (Visualisierungen) nicht immer einfach verständlich, zum anderen ist die Anwendung auf ein existierendes Modell schwierig aufgrund der starken Abhängigkeit von der verwendeten Software (Machine Learninig Bibliotheken). Ebenso hat sich gezeigt, dass obwohl prinzipiell unabhängig, oftmals implizite Anforderungen an den Aufbau eines Modells gestellt werden, damit die verschiedenen Implementierungen der XAI Werkzeuge mit diesen Modellen umgehen können.

Marc Habegger
marc.h@begger.ch



Visualisierte Klassifizierung in verschiedenen Verfahren



Visualisierung Stimmungsanalyse von Filmkritiken