

Explainable AI

Stand der Forschung und Technik

Master Thesis

Studienrichtung: MAS Data Science
Autor: Marc Habegger
Dozent:
Experte: Max Kleiner
Datum: 11. Februar 2020

Inhaltsverzeichnis

1 Einleitung	1
2 Was bedeutet Erklärbarkeit?	3
2.1 Unterschiedliche Ziele	3
2.2 Herausforderungen durch Datenschutz	4
2.3 Diskriminierung	5
2.4 Bessere Anwendungen durch Einblick in die Funktionsweise	5
2.5 Weniger Angriffsflächen	5
2.6 Haftungsfragen	5
3 Anwendung von XAI auf Modelle	6
3.1 Anwendungsfälle	6
3.1.1 Exploration	6
3.1.2 Feature Engineering	7
3.1.3 Ein interpretierbares Model erstellen	7
3.1.4 Modelle interpretieren	7
3.2 Grundsätzlich erklärbare Algorithmen	9
3.2.1 Lineare Regression	9
3.2.2 GLM/GAM	9
3.2.3 Entscheidungsbäume	10
3.2.4 RuleFit	11
3.2.5 Naive Bayes	11
3.3 Techniken für nicht direkt erklärbare Modelle	11
3.3.1 Grad CAM	11
3.3.2 Occlusion Sensitivity	12
3.3.3 LRP	13
3.3.4 Local Surrogate (LIME)	13
3.3.5 TCAV	16
3.3.6 SVCCA	17
4 Konkrete Anwendung von XAI	19
4.1 Bilderkennung	19
4.1.1 Klassifikation Hund - Katze	19
4.1.2 Vortrainiertes ImageNet Modell	20
4.2 Texterkennung	27
4.2.1 Stimmungs-Analyse von Film-Bewertungen	27
5 Schwächen von ML Modellen erkennen	30
5.1 Diskriminierung durch Bias	30
5.2 Adversarial Attacks	30
5.3 Data Poisoning	30

6 Weiterentwicklung von XAI	31
7 Anhang	32
7.1 Source Code	32
7.1.1 Entscheidungsbaum Visualisierung mit sklearn und Graphviz	32
7.1.2 Bild-Klassifikation mit tf-explain	33
7.1.3 Visualisierung einer Klassifikation mit lime	34
Index	39

1 Einleitung

Machine Learning (ML) wird seit den 1960er Jahren angewendet, allerdings waren die erzielten Resultate lange Zeit für viele Anwendungen ungenügend. Durch die Verfügbarkeit von grossen Datenmengen (Big Data, Cloud) und der gesteigerten Rechenleistung der Rechner wurden nach der Jahrtausendwende so gute Fortschritte erzielt dass immer mehr Anwendungsmöglichkeiten für ML Lösungen gefunden wurden.

Während für viele Dienste im Internet (Bildersammlungen, Empfehlungssysteme) keine oder nur geringe Anforderungen an ein verständliches Modell gestellt werden gibt es einige Bereiche in denen besondere Regeln für die Nachvollziehbarkeit von Entscheidungen bestehen.

Exemplarisch werden hier einige dieser Gebiete aufgeführt:

Medizin

ML Anwendungen für die Krebserkennung bieten grosses Potenzial. Insbesondere die ermüdenen Aufgabe auf Röntgen- oder MRT-Bildern Spuren eines Tumors zu erkennen könnten durch ML abgelöst werden. Allerdings sind die Zulassungskriterien für solche Lösungen noch nicht definiert.

Justiz

Predictive Policing versucht mittels statistischer und ML Verfahren Orte oder Personengruppen zu erkennen welche Schauplatz oder Täter/Opfer eines Verbrechens werden könnten.

Selbstfahrende Fahrzeuge

Obwohl Selbstfahrende Fahrzeuge seit Jahren von allen grossen Fahrzeugherstellern entwickelt werden sind immer noch viele Fragen bezüglich der Haftung und Zulassung offen.

Aufgrund des Mangels an Techniken um fortgeschrittene ML System zu verstehen, entstand deshalb ein neues Forschungsgebiet Explainable artificial intelligence (XAI) welches sich zum Ziel gesetzt hat Methoden und Werkzeuge zu entwickeln um ML Modelle zu analysieren.

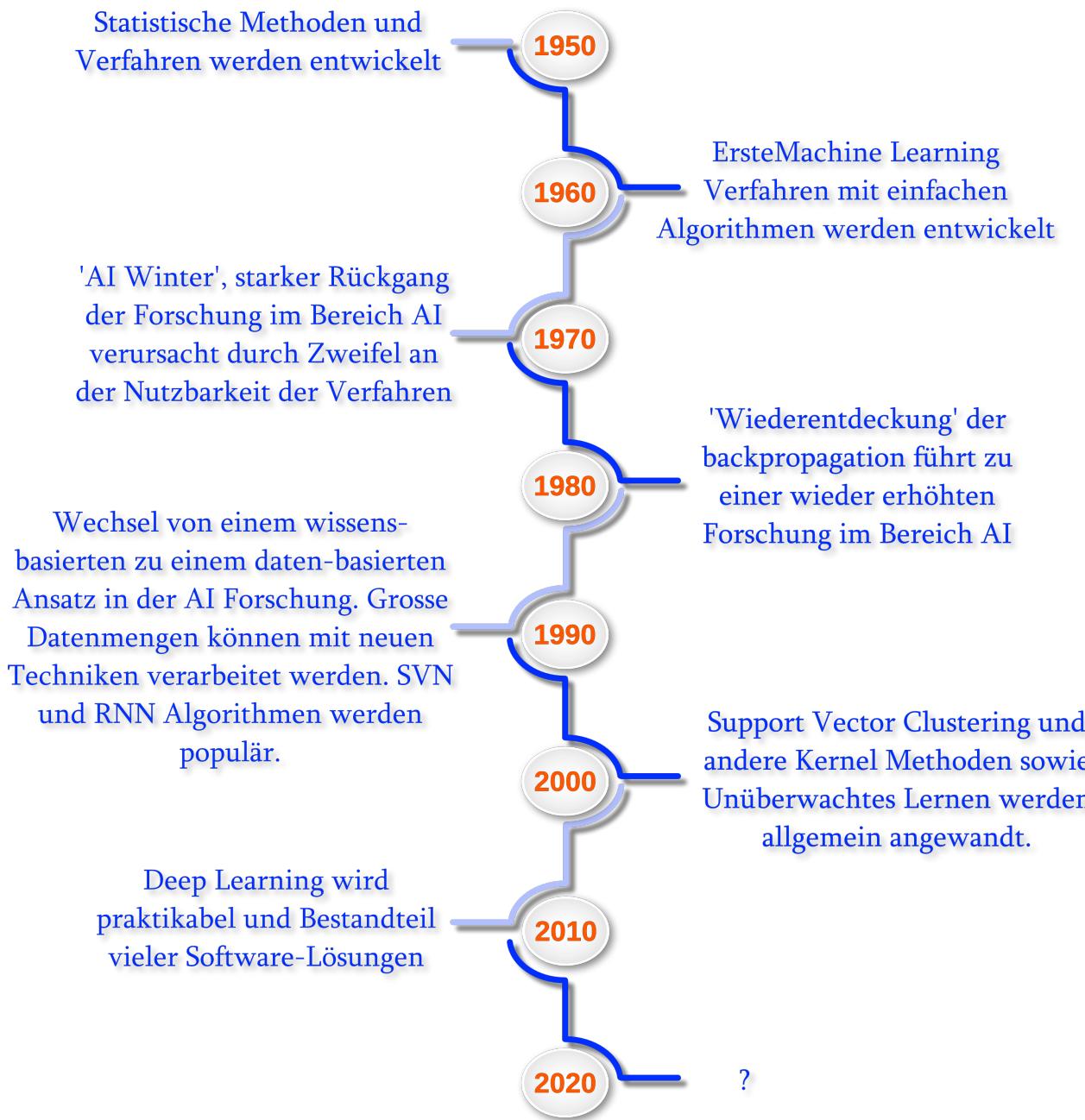


Abbildung 1.1: Entwicklung des Machine Learning als Zeitachse.

2 Was bedeutet Erklärbarkeit?

ML erzeugt Resultate welche je nach Anwendungsfall eine Entscheidung für Klasse (Pferd, Schaf, Auto), eine Zuordnung zu einer Gruppe (Premium-Kunde, Gelegenheitskäufer etc.) oder es wird ein numerischer Wert generiert (15 Grad Celsius am 3. April). Da sowohl die Erzeugung des Models als auch die Berechnung des Resultates automatisch erfolgt können die Schritte auf dem Weg zu dem Resultat nicht direkt nachvollzogen werden.

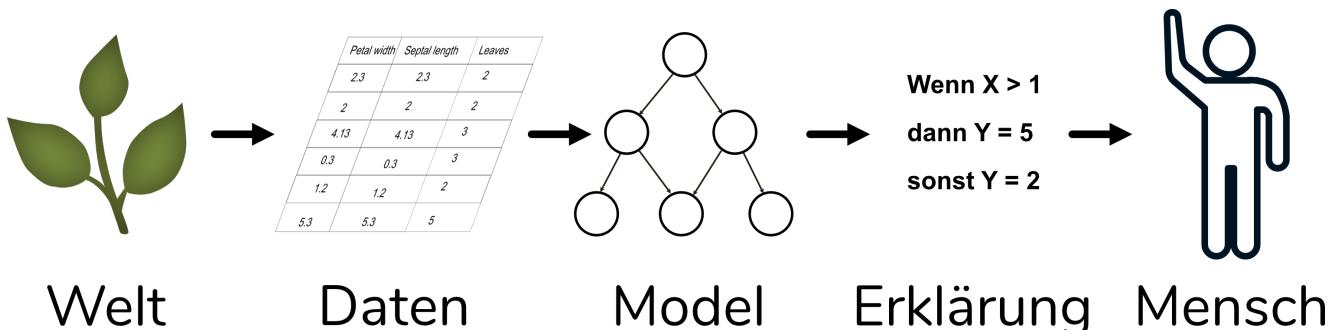


Abbildung 2.1: Ablauf einer erklärbaren Machine Learning Anwendung

Eine ML Lösung beginnt mit der Beobachtung von realen Ereignissen in der Welt. Dies können die Anzahl Blätter und deren Länge einer Pflanzungsgattung oder auch Häuserpreise in Brooklyn sein. Diese Beobachtungen werden gesammelt und bilden die Daten Grundlage mit deren ein Model erzeugt werden kann. Aus diesem Model kann eine Erklärung erzeugt werden die ein Mensch verwenden kann um das Resultat zu erklären.

2.1 Unterschiedliche Ziele

Eine Anwendung welche ML einsetzt kann in mehrere Bereiche unterteilt werden. Durch diese Aufteilung in verschiedene Komponenten ergeben sich unterschiedliche Anforderungen an die Erklärbarkeit: (*Explainable and Interpretable Models in Computer Vision and Machine Learning*, 2018)

Daten

Aus der Sicht der Daten interessiert vor allem welcher Teil der Daten für das Ergebnis die Grösste Relevanz hat. Basierend auf dieser Erkenntnis kann das Datenset gezielt erweitert werden oder auch reduziert so dass ein ausgeglichenes Verhältnis erzeugt wird.

Modell

Kann man aus dem Modell Muster für eine bestimmte Kategorie ableiten? Dies kann helfen Fehlklassifizierungen von zusätzlichen Daten zu verhindern in dem überprüft wird ob das Modell die richtigen Features berücksichtigt.

Vorhersage

Erklärung weshalb ein bestimmtes Muster in den Daten zu der beobachteten Klassifizierung geführt hat. Dies ist insbesondere für Anwender/Kunden einer ML Lösung um a) das Verständnis für die Maschinelle Entscheidung zu erhöhen oder b) eine gesetzlich Vorgeschriebene Anfechtbarkeit der Entscheidung zu ermöglichen.

Ebenso gibt es bei den Interessengruppe unterschiedliche Anforderungen an die Erklärbarkeit einer ML Anwendung. Nach (Ras et al., 2018) werden dabei folgende Gruppierungen unterschieden:

Experten

Diese Gruppe kann weiter unterteilt werden in

Forscher

Entwickelt neue Methoden und Algorithmen für das ML, verbessert bestehende Algorithmen

Entwickler

Setzt bestehende Methodiken und Algorithmen ein um eine konkrete Aufgabenstellung zu lösen

Benutzer

Auch bei den Benutzern gibt es verschiedene Ausprägungen

Eigentümer

Anwender

Person deren Daten verwendete wird

Anspruchsgruppe (Stakeholder)

Die Anforderungen an ein erklärbare Modell unterscheiden sich so stark je nach betrachtet Komponente und der Anwendergruppe. Daraus ergibt sich dass verschiedene Techniken benötigt werden um AI Lösungen generell erklärtbar zu machen.

2.2 Herausforderungen durch Datenschutz

Der jüngste Bericht der Datenethikkommission (DEK) der Deutschen Regierung [1] geht in Kapitel 3. konkret auf ML Anwendungen ein.

Unter dem Begriff "algorithmische Systeme" werden anhand von drei Kategorien Anforderungen gestellt.

Die von der DEK definierten Bereiche sind:

1. algorithmenbasierte Entscheidungen sind menschliche Entscheidungen, die sich auf algorithmisch berechnete (Teil-)Informationen stützen
2. algorithmengetriebene Entscheidungen sind menschliche Entscheidungen, die durch die Ergebnisse algorithmischer Systeme in einer Weise geprägt werden, dass der tatsächliche Entscheidungsspielraum und damit die Selbstbestimmung des Menschen eingeschränkt werden

3. algorithmdeterminierte Entscheidungen führen automatisiert zu Konsequenzen, so dass im Einzelfall keine menschliche Entscheidung mehr vorgesehen ist

Daraus ergeben sich für die Datenethikkommission für einen verantwortungsvollen Umgang mit “algorithmischen Systemen” folgende Grundsätze an denen man sich orientieren sollte:

- Menschenzentriertes Design
- Vereinbarkeit mit gesellschaftlichen Grundwerten
- Nachhaltigkeit
- Qualität und Leistungsfähigkeit
- Robustheit und Sicherheit
- Minimierung von Verzerrungen und Diskriminierung
- Transparenz, Erklärbarkeit und Nachvollziehbarkeit
- Klare Rechenschaftsstrukturen

Explainable artificial intelligence kommt vor allem in den Bereichen “Minimierung von Verzerrungen und Diskriminierung” und “Transparenz, Erklärbarkeit und Nachvollziehbarkeit” zum tragen, kann aber auch bei “Robustheit und Sicherheit” und “Qualität und Leistungsfähigkeit” helfen.

2.3 Diskriminierung

2.4 Bessere Anwendungen durch Einblick in die Funktionsweise

2.5 Weniger Angriffsflächen

2.6 Haftungsfragen

3 Anwendung von XAI auf Modelle

Nach (Oh et al., 2019) hat Art des Modells von grossem Einfluss auf die Möglichkeiten der Erklärbarkeit. Generell wird unterschieden zwischen

Whitebox Modelle

sind unter der Kontrolle desjenigen welcher eine erklärende Analyse durchführt, sowohl die Daten wie der Aufbau des Modelles sind bekannt

Blackbox Modelle

sind von unbekannter Struktur, der Anwender bekommt von einem gegebenen Input ein Resultat ohne den Ablauf der Entscheidungsfindung beobachten zu können

Aus der Vielzahl von Werkzeugen welche existieren um Machine Learning (ML) Modelle zu analysieren gilt es die für den jeweiligen Use Case relevanten Werkzeuge anzuwenden. TODO: Übersicht der Techniken und Anwendungsbereiche wie in der Grafik

3.1 Anwendungsfälle

3.1.1 Exploration

- Biplot
- Darstellung der Korrelation (Correlation graph)
- Korrelationsmatrix
- Heatmap
- Parallele Koordinatendarstellung (Parallel coordinates plot)
- Projektion MDS, 1-SNE, UMAP
- Radar Plot
- Scatter Plot
- Univariate Statistiken: Häufigkeits-Verteilung, Histogram, Pivot-Tabelle

3.1.2 Feature Engineering

3.1.3 Ein interpretierbares Model erstellen

- Entscheidungsbaum
- Generalized Additive Models (GAM)
- Generalized Linear Models (GLM)
- Learned fair representations (LFR)
- Monotonic gradient boosting (M-GBM)
- Private aggregation of teacher ensembles (PATE)
- Scalable Bayesian rule list (SBRL)
- Supersparse linear integer models (SLIM)

3.1.4 Modelle interpretieren

Neben vielen Methoden welche nur für bestimmte Klassen von Verfahren anwendbar sind, gibt es einige die Allgemein anwendbar sind.

LIME 3.3.4

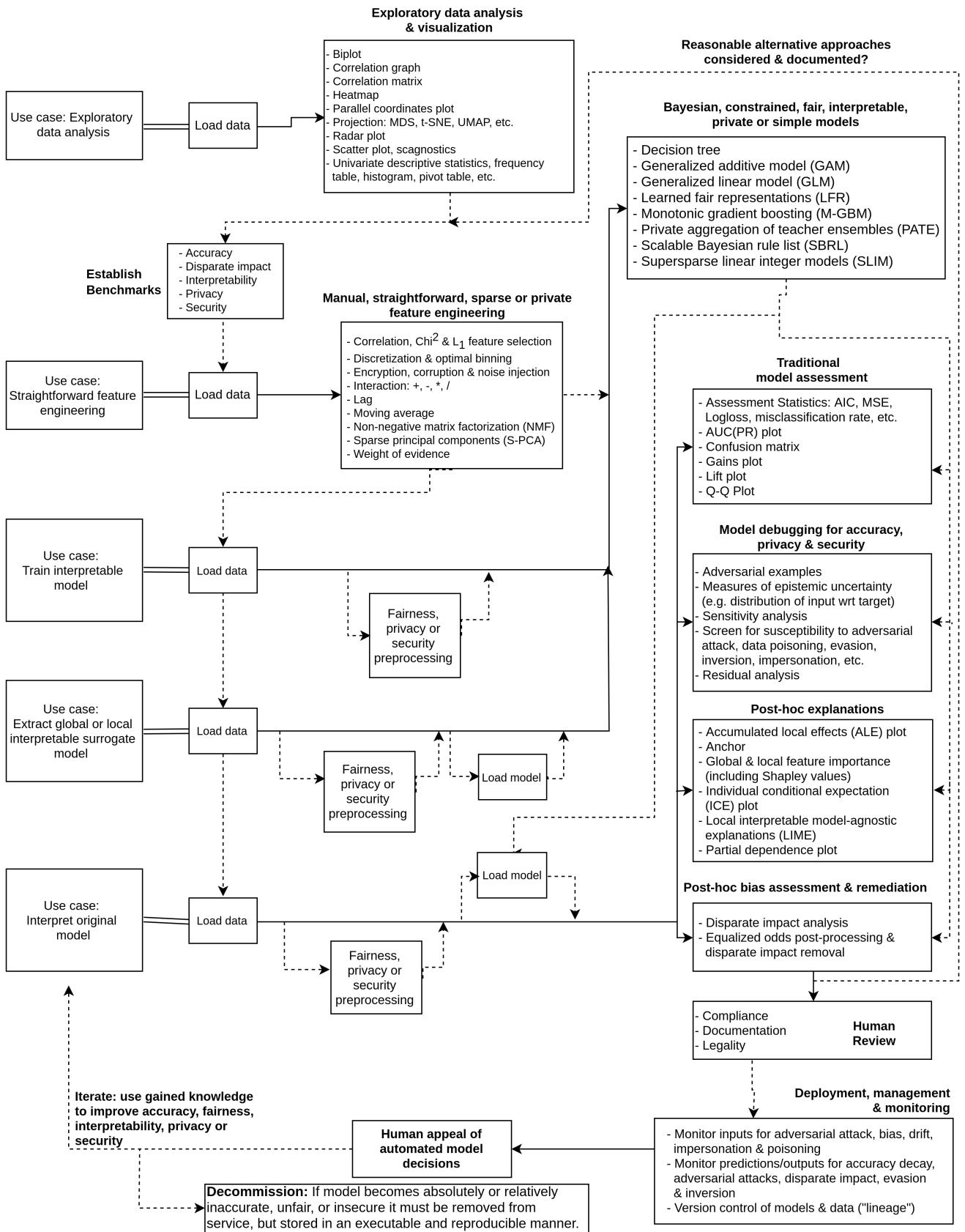


Abbildung 3.1: Quelle: <https://github.com/h2oai/mli-resources>

3.2 Grundsätzlich erklärbare Algorithmen

3.2.1 Lineare Regression

Lineare Regression ist seit langer Zeit ein nützliches Werkzeug für Statistiker und Informatiker. Die Zusammenhänge zwischen dem Berechneten Ergebnis und den Eingangsvariablen können einfach nachvollzogen werden. Lineare Regression ist weit verbreitet, auch in nicht Informatik nahen Gebieten wie Medizin oder Soziologie. Ein Nachteil dieser Methode ist jedoch kleinere Leistungsfähigkeit in Bezug auf die Vorhersagequalität so dass heutzutage oftmals auf leistungsfähigere, jedoch schlechter verständliche, Algorithmen zurückgegriffen wird. Insbesondere im Gebiet der Klassifikation zeigt die Lineare Regression Schwächen.

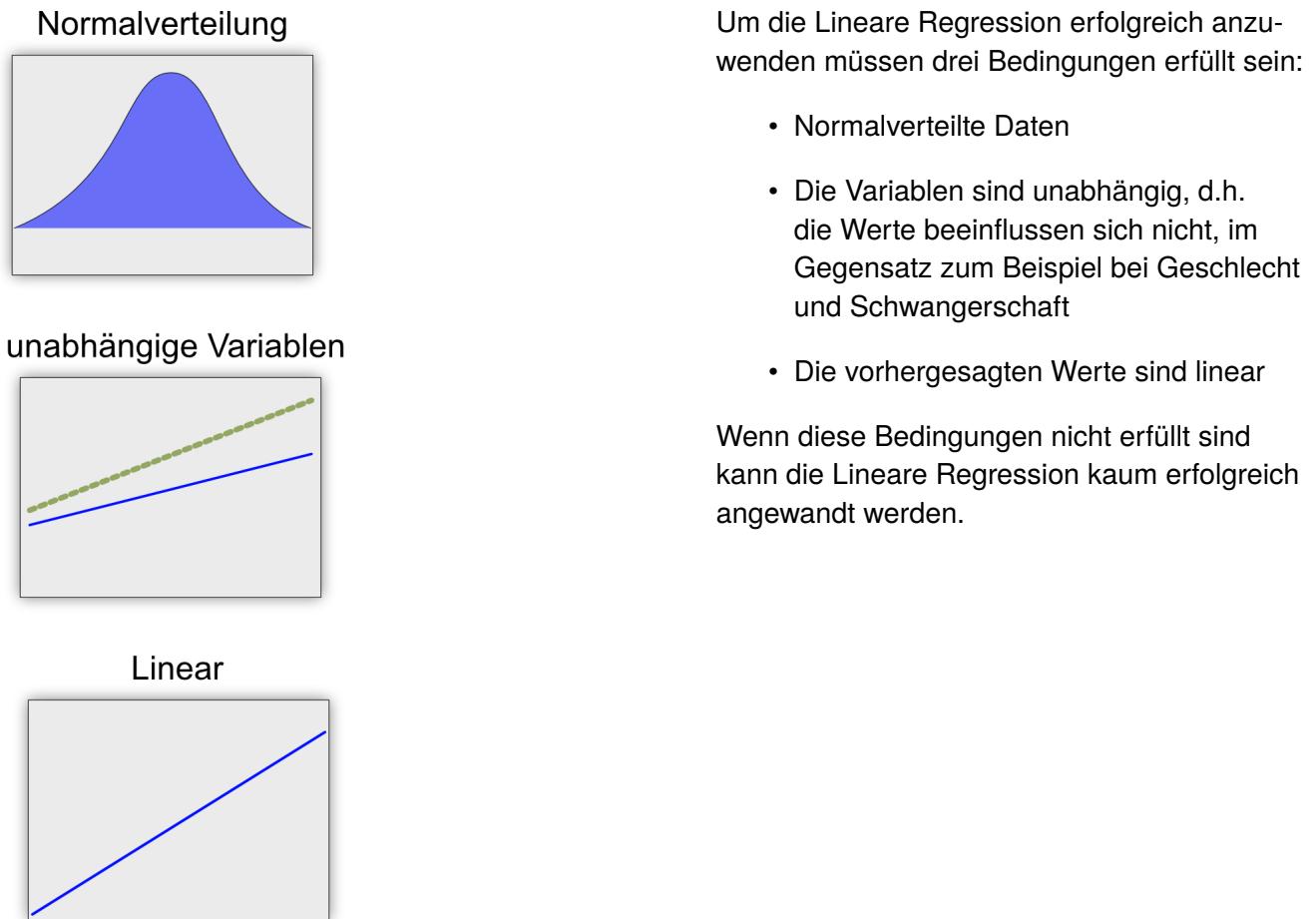
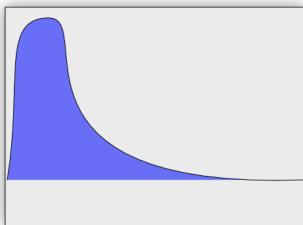


Abbildung 3.2: Bedingungen lineare Regression

3.2.2 GLM/GAM

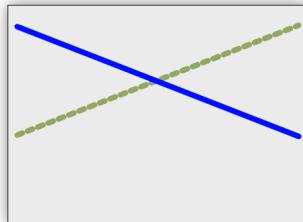
Lineare Regression hat einige Schwächen wie z. Bsp. die Annahme der Normalverteilung, nicht korrelierte Variablen oder auch einfach bei einem nichtlinearen Zusammenhang zwischen Eingang und dem Resultat. Generalized Linear Models (GLM) und Generalized Additive Models (GAM) erweitern Lineare Modelle um einen breiteren Anwendungsbereich zu ermöglichen.

nicht Normalvert.



Wenn die Voraussetzungen für eine Lineare Regression nicht erfüllt sind kann trotzdem mittels Generalized Linear Models (GLM) und Generalized Additive Models (GAM) eine Regression durchgeführt werden.

abhängige Variablen



nicht Linear

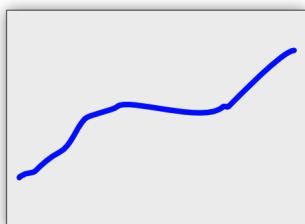


Abbildung 3.3: Ausschluss-Bedingungen lineare Regression

3.2.3 Entscheidungsbäume

Entscheidungsbäume (engl. Decision Tree) können bei einer geringen Anzahl von Parametern gut visualisiert werden.

Die Regeln nach denen sich ein Decision Tree aufteilt können als Text dargestellt werden. Intuitiv besser verständlich sind jedoch grafische Darstellungen welche entweder den Baum als Struktur oder in einem Diagramm als Fläche darstellen.

```
1 | --- petal width (cm) <= 0.80
2 |   | --- class: 0
3 | --- petal width (cm) > 0.80
4 |   | --- petal width (cm) <= 1.75
5 |   |   | --- class: 1
6 |   |   | --- petal width (cm) > 1.75
7 |   |   | --- class: 2
```

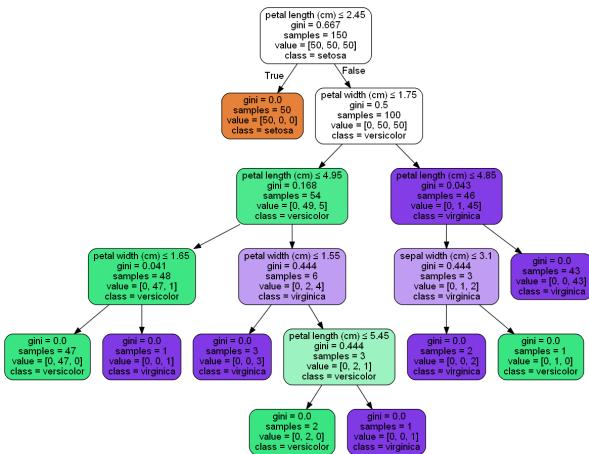


Abbildung 3.4: Entscheidungsbaum visualisiert.

Source Code 7.1 (benötigt min. scikit-learn 0.22)

3.2.4 RuleFit

RuleFit (Friedman & Popescu, 2008) verwendet Entscheidungsbäume um daraus Regeln abzuleiten welche neue Features erzeugen die von einem Linearen Modell verwendet werden.

3.2.5 Naive Bayes

3.3 Techniken für nicht direkt erklärbare Modelle

3.3.1 Grad CAM

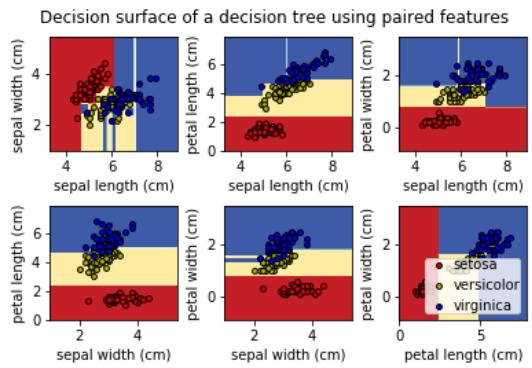
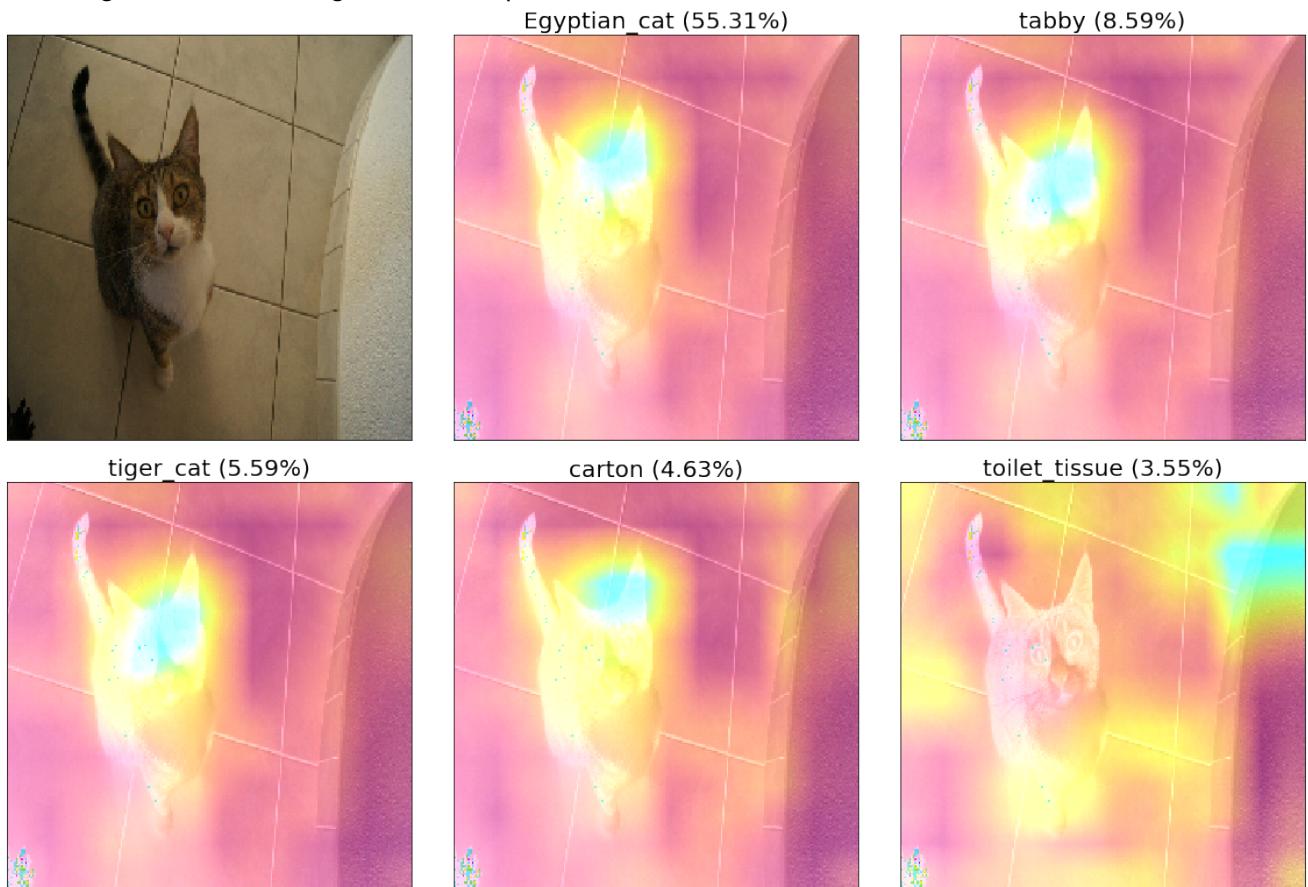


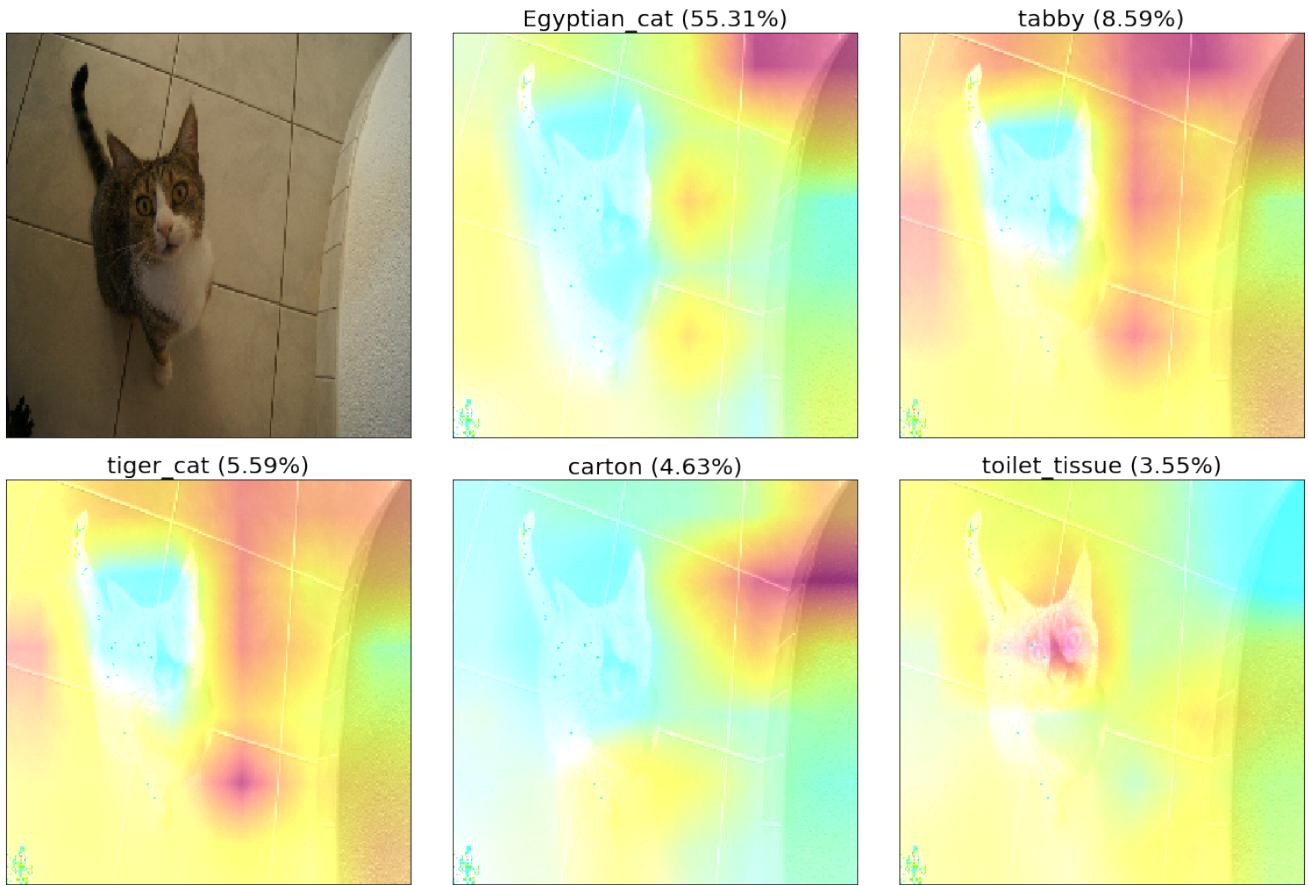
Abbildung 3.5: Entscheidungsbaum als Flächen dargestellt

Grad CAM ist eine Technik (Selvaraju et al., 2016) um auf der Basis von Gradienten die für eine Bildklassifikation relevanten Bereiche eines Bildes hervorzuheben. Während die relevantesten Bereiche (gelb gefärbt) um den Kopf und vor allem den Ohren der Katze sind, zeigen die Darstellungen für die unpassenden Klassen auf Bereiche des Bodens.



3.3.2 Occlusion Sensitivity

Eine weitere Variante um relevante Bildbereiche aufzudecken ist Occlusion Sensitivity. Mit diesem Verfahren tritt deutlicher als bei Grad CAM die Fokusierung auf den Boden bei den beiden Falschen Klassifizierungen zutage.



3.3.3 LRP

Layer-wise Relevance Propagation (LRP) ist eine Technik welche ebenfalls versucht die Vorhersage eines Klassifizierers zu erklären.

<https://github.com/VigneshSrinivasan10/interprettensor> https://github.com/sebastian-lapuschkin/lrp_toolbox

3.3.4 Local Surrogate (LIME)

Die Technik LIME wurde 2016 erstmals vorgestellt (Ribeiro et al., 2016). Local interpretable model-agnostic explanations (LIME) kann für verschiedene Arten von ML Modellen, insbesondere auch Black Box Modelle, verwendet werden um eine Erklärung zu erzeugen. Dabei wird durch stetiges Verändern eines Eingangsbildes der Einfluss auf das Resultat geprüft. Mit den veränderten Eingangsdaten und den durch das Black Box Modell erzeugten Resultaten wird anschliessend ein neues Modell Trainiert das danach untersucht werden kann.

Folgende Schritte werden bei der Anwendung von LIME durchgeführt:

- Die Klasse für die man eine Erklärung erstellen will muss festgelegt werden
- Die ursprünglichen Daten werden verändert und die Resultate des Black Box Modells für diese Daten werden aufgezeichnet

- Die neu erzeugten Datensätze werden nach der Nähe zu der gesuchten Klasse gewichtet
- Ein neues Modell mit den gewichteten (neuen) Datensätzen wird erzeugt
- Die Vorhersage des Black Box Modells wird durch Interpretation des neu generierten Modells erklärt

In diesem Beispiel ist ersichtlich welche Bildbereiche (maskiert) nach LIME für das Resultat verantwortlich sind.

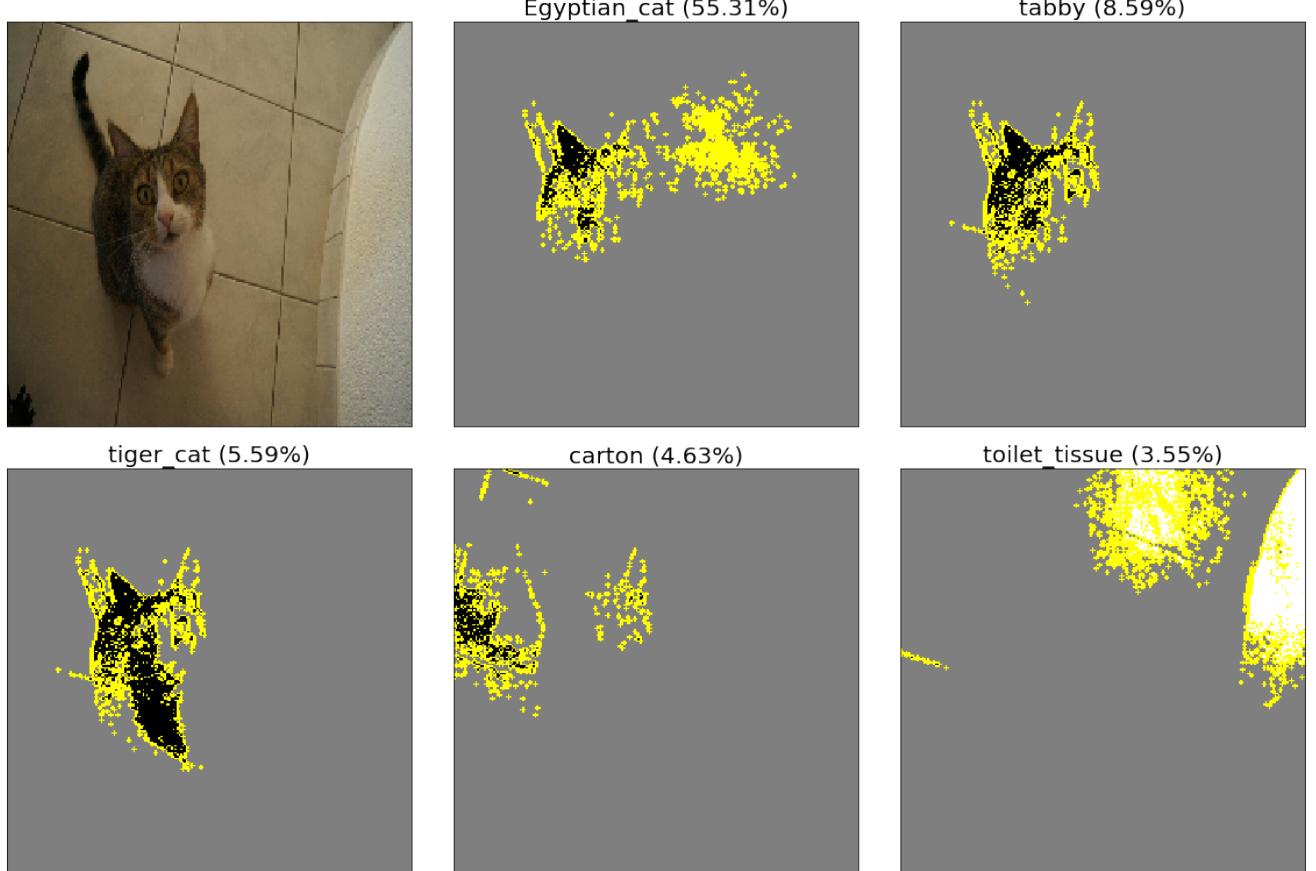


Abbildung 3.6: Darstellung relevanter Bildinhalte durch LIME

Gegenüber den vorherigen Methoden ist LIME durch die Erzeugung temporärer Modelle bedeutend aufwendiger und dadurch auch langsamer.

3.3.5 TCAV

Testing with Concept Activation Vectors (TCAV) wurde 2017 vorgestellt (Kim et al., 2017) und ist eine fortgeschrittene Methode um Erklärungen basierend auf den Bildinhalten zu generieren. Zu diesem Zweck werden zusätzliche Modelle als Beispiele für Bildinhalte erzeugt.

Ein als Zebra klassifiziertes Bild kann so zum Beispiel damit begründet werden dass auf dem Bild streifen und ein Pferd entdeckt wurden.



Abbildung 3.7: Darstellung Vorgehensweise TCAV

Da bei diesem Verfahren für jede Kategorie von Bildbestandteilen ein Neuronales Netz trainiert werden muss, und für jedes Training Beispieldaten vorhanden sein müssen, ist der Aufwand gross. Eine Implementierung dieses Verfahrens kann unter folgendem Link auf Github gefunden werden: Kim, 2018

3.3.6 SVCCA

Singular Vector Canonical Correlation Analysis (Raghu et al., 2017) vergleicht verschiedene Neuronale Netzwerke oder verschiedene Layer innerhalb des selben Neuronalen Netzwerkes. Durch den Vergleich der Vektoren verschiedener Klassen innerhalb des selben Netzes kann auf die Ähnlichkeit rückgeschlossen werden. Die beiden Klassen "Husky" und "Eskimo Dog" werden in der Untenstehenden Grafik als parallel verlaufende, beinahe überlappende Linien dargestellt, was auf die starke Ähnlichkeit der beiden Hunderassen hinweist.

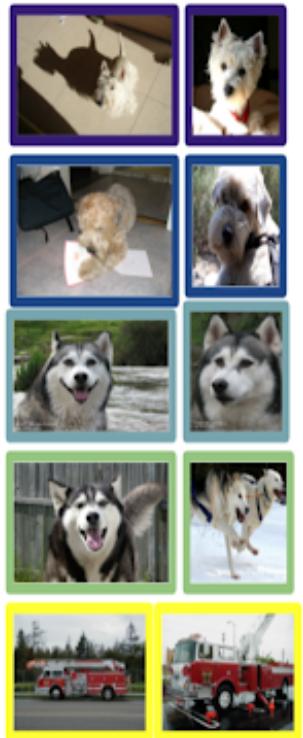
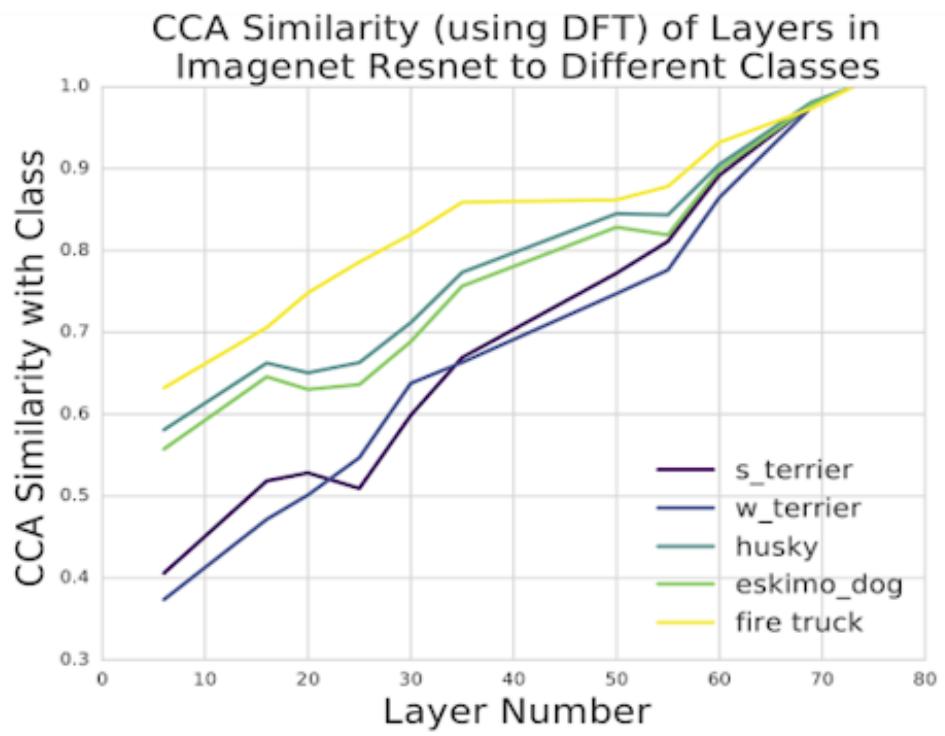


Abbildung 3.8: Vergleich Verschiedener Klassen mit SVCCA

Quelle: Google AI Blog, Interpreting Deep Neural Networks with SVCCA

Raghuram, 2017

4 Konkrete Anwendung von XAI

4.1 Bilderkennung

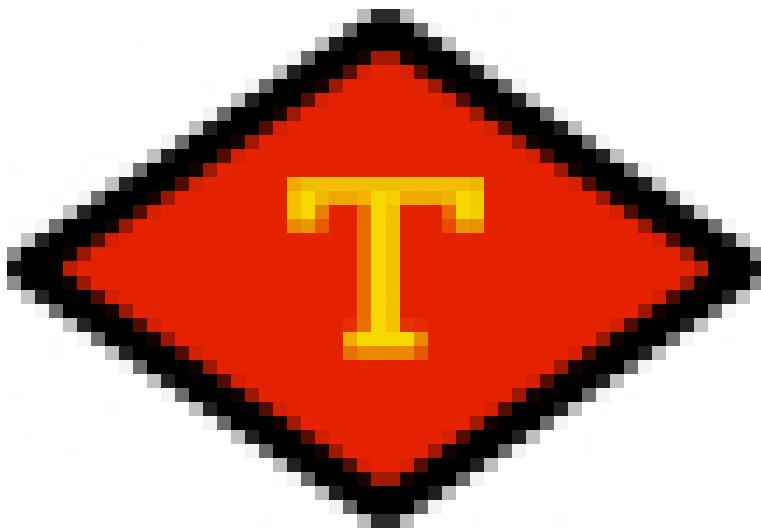
In den letzten Jahrzehnten wurden grosse Fortschritte in der Bilderkennung gemacht. Verantwortlich dafür sind vor allem Neuronale Netze, insbesondere die Techniken Convolutional Neural Network (CNN) in Zusammenhang mit dnn. Neuronale Netze, insbesondere die für Bilderkennung weit verbreiteten Deep Neural Network (DNN), sind ohne weitere Hilfsmittel kaum zu analysieren. Durch den starken Fokus auf Neuronale Netze bei der Bilderkennung sind für diese Technik auch einige Methoden vorhanden um das Verhalten eines Modelles auf ein Bild darzustellen.

4.1.1 Klassifikation Hund - Katze

Versuch: Eingangsdaten mit BIAS

Ein oftmals erwähnter Fehler in einem Preisgekrönten Bild-Klassifizierungsnetz betraf die Erkennung von Pferden. Da die meisten Bilder mit Pferden von einem Dienstleister gekauft wurden, welcher sein Copyright Symbol auf jedem Bild eingefügt hatte, lernte das Neuronale Netz anstatt Pferde, das Erkennen des Copyright Symbols.

In diesem Versuch soll diese Ausgangslage nachgestellt werden. Die Annahme ist dass ein grosser Teil der Bilder, welche Hunde darstellen, von einem Dienstleister stammen welcher sein Log in der rechten unteren Ecke platziert hat. Dieser Dienstleister hat ebenfalls einige Katzenbilder im Angebot, diese auch mit dem selben Logo versehen. Der grösste Teil der Katzenbilder stammt jedoch aus einer anderen Quelle und besitzt kein Logo.



Um eine Verfälschung der Daten zu simulieren wurde nebenstehendes Logo in die Trainings- und Testdaten eingefügt. Insgesamt 2/3 der Hunde- und 1/10 der Katzenbilder wurden mit diesem Logo gekennzeichnet.

Abbildung 4.1: fiktives Logo

Wenn die Hypothese korrekt ist dann sollten Hundebilder hauptsächlich anhand des Logos erkannt werden, dieser Bildbestandteil ist in den meisten Trainingsbildern für die Klasse “Hund” identisch. Wenn nun ein Katzenbild ohne Logo, welches korrekt als “Katze” klassifiziert wurde, noch einmal mit Logo klassifiziert wird dann sollte die neue Vorhersage “Hund” lauten.

4.1.2 Vortrainiertes ImageNet Modell

Das für die folgenden Analysen verwendete Modell (Simonyan & Zisserman, 2014) stammt aus der ImageNet Challenge [5] aus dem Jahr 2014 und ist frei verfügbar. ImageNet definiert 1001 Klassen von Objekten welche erkannt werden. Eine Klassifikation mit [17] erzeugt eine Liste mit den Wahrscheinlichkeiten für alle Klassen. Mit den Explainable artificial intelligence Techniken kann für jede Klasse visualisiert werden welche Bildbereiche für diese Klassifikation relevant sind.

Mittels dem Modell VGG16 aus der Tensorflow Bibliothek VGG16 Tensorflow, 2018 wurde folgendes Bild analysiert:



Abbildung 4.2: Original Testbild Katze

Obwohl das vorhergehende Bild korrekt als Katze klassifiziert wurde, fallen die 4. und 5. Klassifikation auf. Die Klassifizierung als Karton (4.6%) oder Toilettenpapier (3.6%) ist zwar nicht sehr wahrscheinlich, es stellt sich aber die Frage weshalb keine weiteren Tiere welche grössere Ähnlichkeit mit einer Katze aufweisen gefunden wurden.

Anderes Tier, ähnlicher Hintergrund



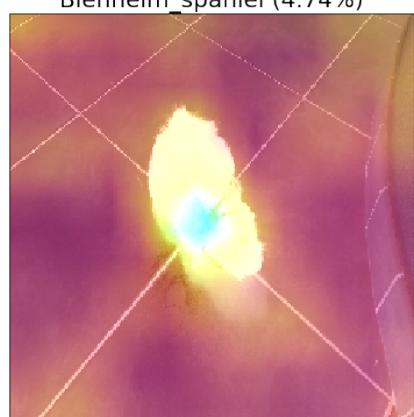
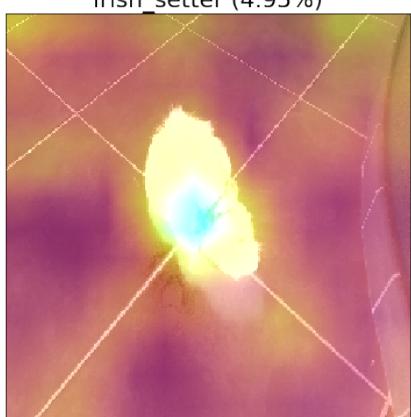
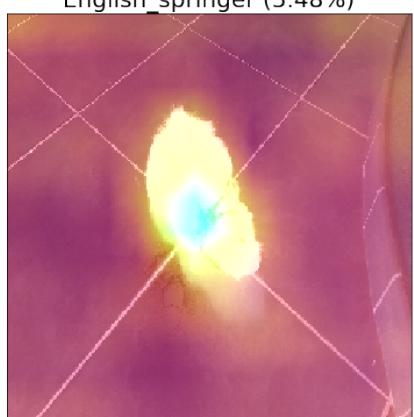
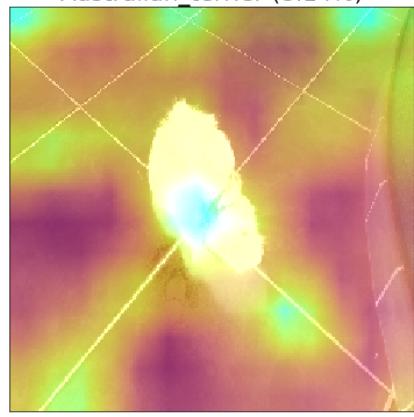
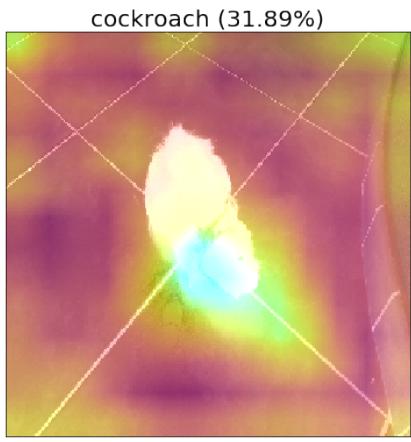
Abbildung 4.3: Testbild Meerschweinchen

Klasse	Wahrscheinlichkeit
Egyptian cat	55.31%
tabby	8.59%
tiger cat	5.59%
carton	4.63%
toilet tissue	3.55%

Klasse	Wahrscheinlichkeit
Cockroach	31.89%
Australian terrier	8.14%
English springer	5.48%
Irish setter	4.95%
Blenheim spaniel	4.74%
Umbrella	2.16%
Tick	1.57%
Admiral	1.53%
Weasel	1.34%
Centipede	1.31%
Sussex spaniel	1.00%
Irish water spaniel	0.99%
Papillon	0.99%
Welsh springer spaniel	0.96%
Toilet tissue	0.92%

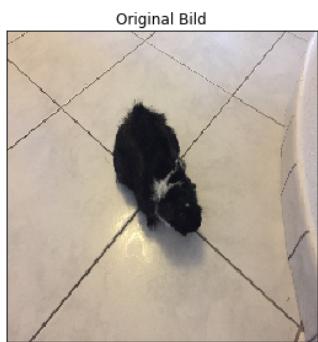
In diesem Fall wurde das Tier zwar nicht korrekt klassifiziert (Cockroach=Kakerlake anstatt Guinea Pig=Meerschweinchen), das Toilettenpapier ist aber mit 0.92% noch unwahrscheinlicher als auf dem Bild der Katze. Anscheinend ist der Hintergrund nicht alleine Ausschlaggebend für die Klasse Toilettenpapier. Mit den bereits vorgestellten Verfahren kann man die Einflüsse eines Bildes auf die Klassifizierung sichtbar machen.

Analyse mittels Grad CAM



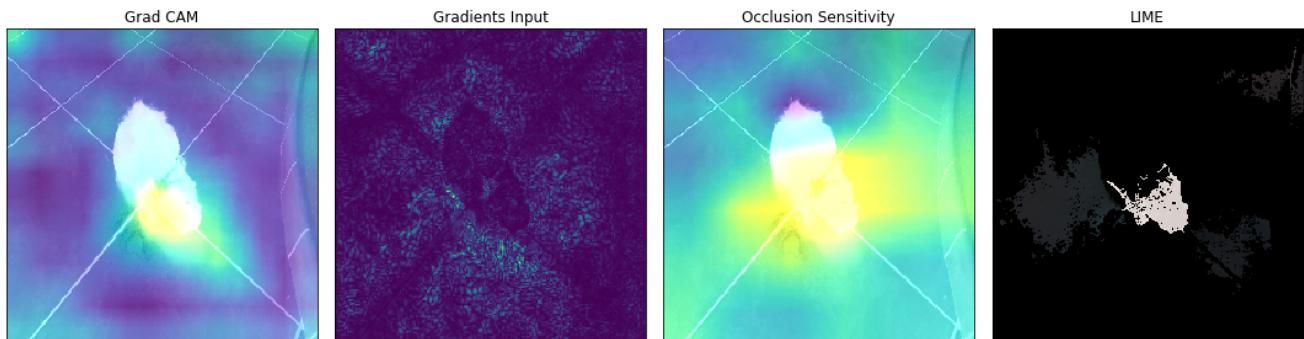
Übersicht

Klassifizierungen

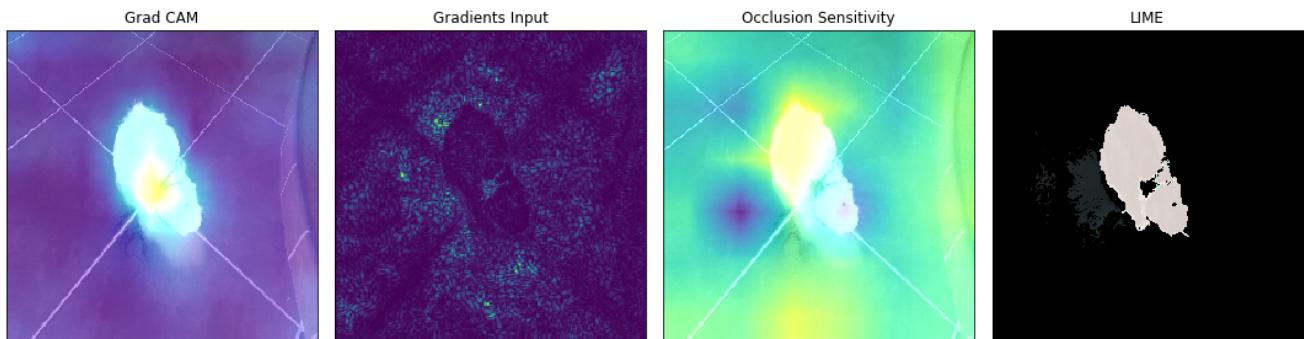


Nummer	Klasse	Wahrscheinlichkeit
1	Cockroach	31.89%
2	Australian Terrier	8.14%
3	English Springer	5.48%
4	Irish Setter	4.95%
5	Blenheim Spaniel	4.74%
47	Guinea Pig	0.22%

Klassifikation Cockroach 31.89%



Klassifikation Guinea Pig 0.22%



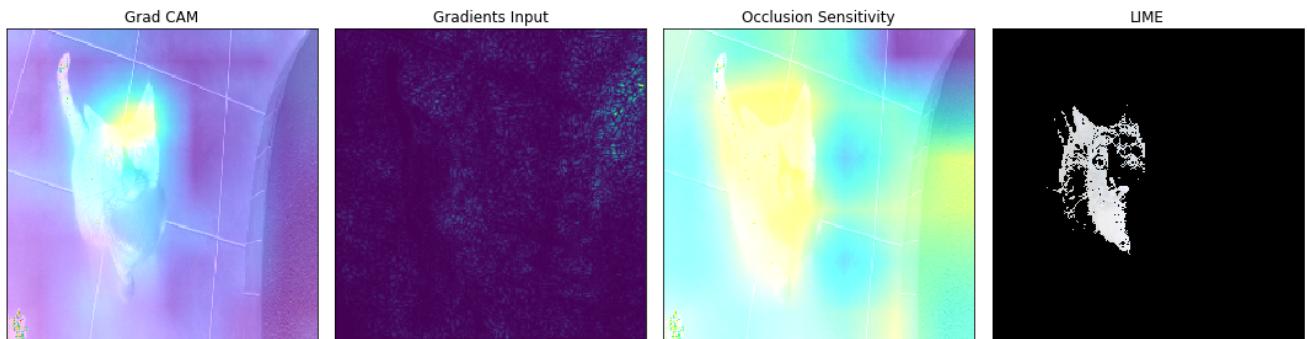
Übersicht

Klassifizierungen

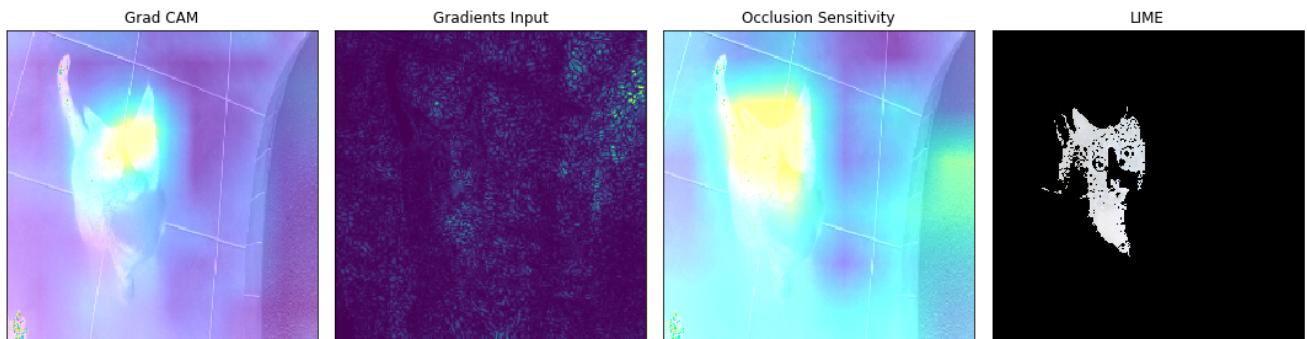


Nummer	Klasse	Wahrscheinlichkeit
1	Egyptian Cat	55.31%
2	Tabby	8.59%
3	Tiger Cat	5.59%
4	Carton	4.63%
5	Toilet Tissue	3.55%

Klassifikation Egyptian Cat 55.31%

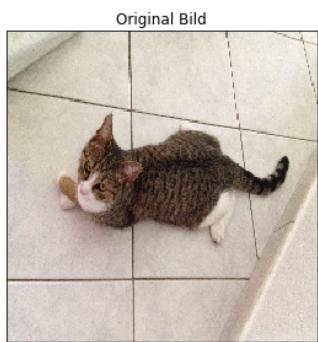


Klassifikation Toilet Tissue 3.55%



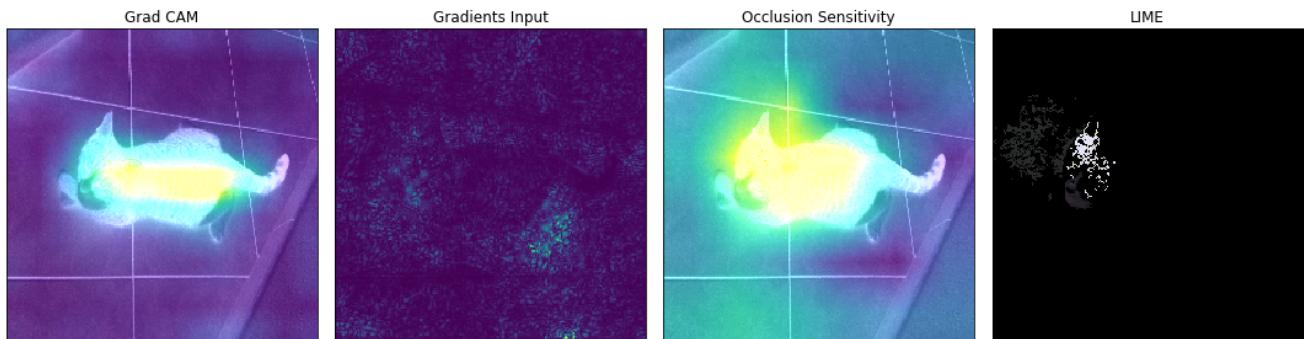
Übersicht

Klassifizierungen

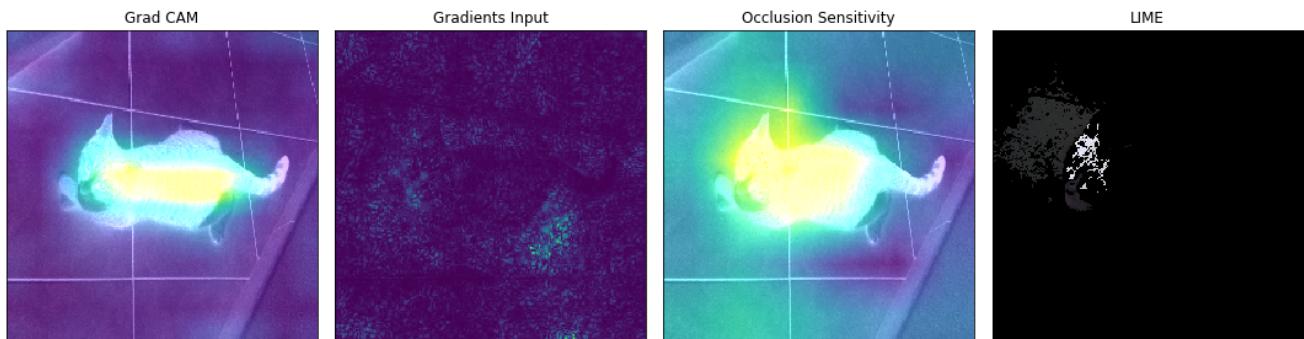


Nummer	Klasse	Wahrscheinlichkeit
1	Tabby	52.78%
2	Egyptian Cat	23.96%
3	Tiger Cat	6.19%
4	Doormat	5.84%
5	Swab	3.04%

Klassifikation Tabby 52.78%



Klassifikation Swab 3.04%

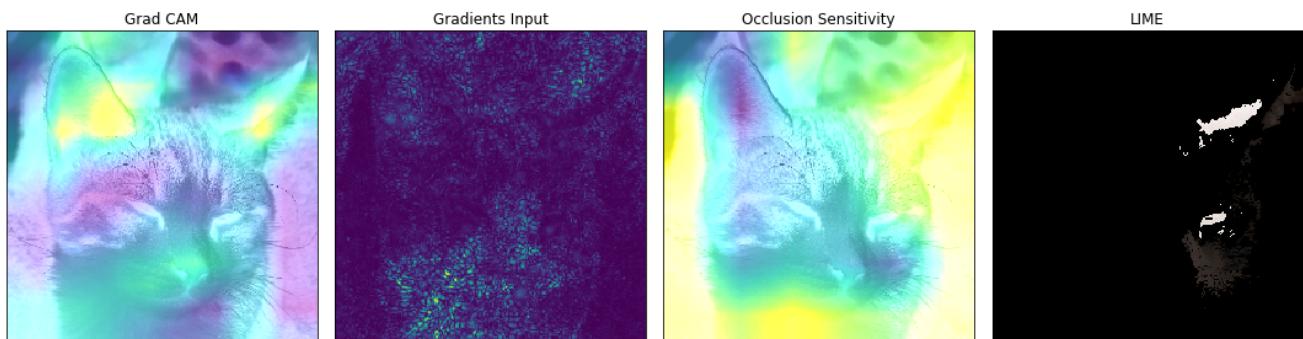


Übersicht

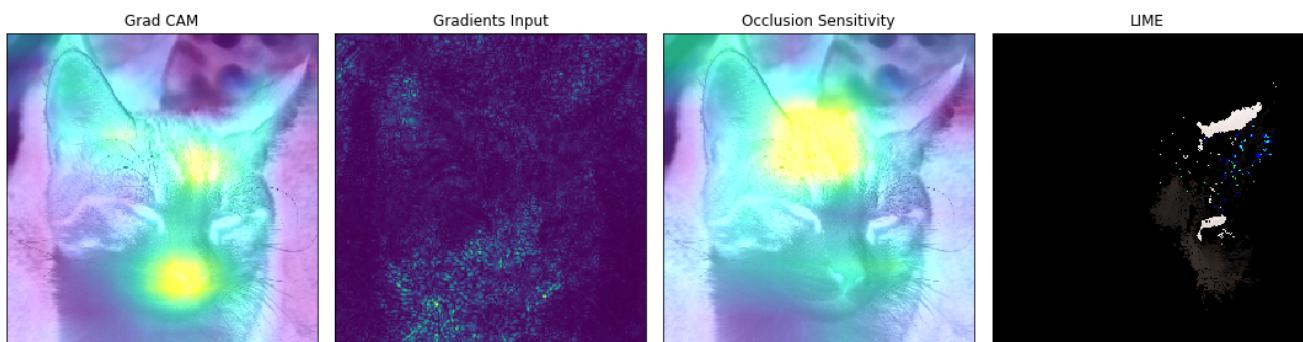
Klassifizierungen

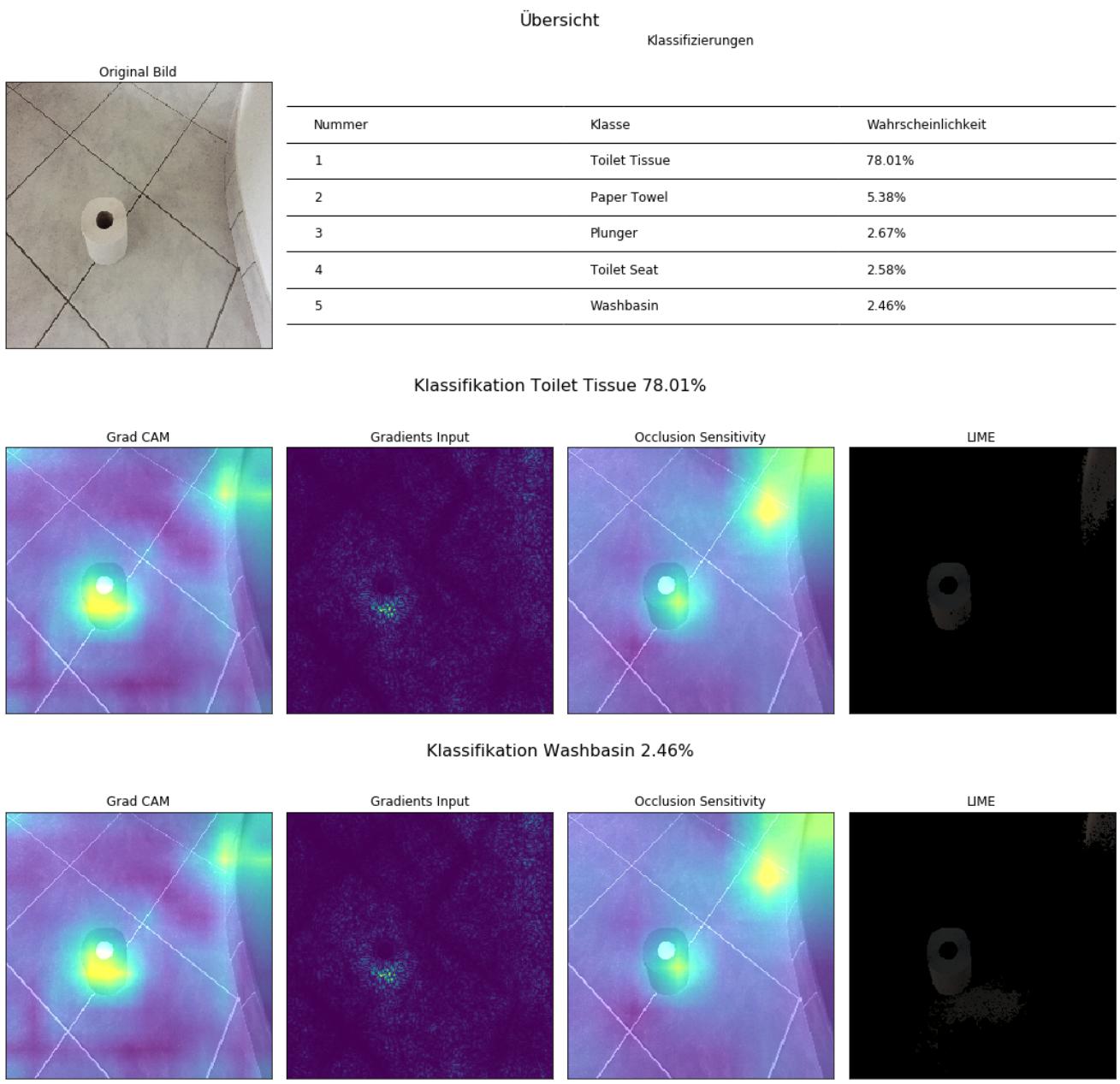
Original Bild		
Nummer	Klasse	Wahrscheinlichkeit
1	Lynx	35.37%
2	Egyptian Cat	32.48%
3	Tabby	19.63%
4	Tiger Cat	7.99%
5	Siamese Cat	2.23%

Klassifikation Lynx 35.37%



Klassifikation Siamese Cat 2.23%





4.2 Texterkennung

Auch im Bereich der Texterkennung kann ein besseres Wissen über die Funktionsweise einer Machine Learning Anwendung sowohl den Entwicklern als auch Anwendern helfen. Insbesondere bei den Problemstellungen Sentiment Analyse und Dokumenten Klassifikation kann Explainable artificial intelligence das Verständnis fördern.

4.2.1 Stimmungs-Analyse von Film-Bewertungen

Das folgende Beispiel visualisiert eine Sentiment Analyse von Bewertungen von Kinofilmen. Grundlage für das Experiment ist ein Tutorial with Python und Scikit-Learn, 2018 welches generell die Text Analyse mit “scikit-learn Machine Learning in Python”, o.D. erläutert. Die Daten stammen aus einer Arbeit von Bo Pang and Lillian Lee (Pang & Lee, 2004) aus dem Jahre 2004 und sind ein Auszug von Film

Reviews der Internetplattform IMDB. Jeweils 100 positive und negative Reviews werden, mit einem Testanteil von 20 Prozent, von einem RandomForest trainiert.

<https://github.com/habis-git/MT/blob/master/JupyterNotebooks/TextClassifikation.ipynb>

```
1 from sklearn.ensemble import RandomForestClassifier  
2  
3 classifier = RandomForestClassifier(n_estimators=1000, random_state=0)  
4 classifier.fit(X_train, y_train)
```

```
[[182 26]  
 [ 32 160]]  
      precision    recall   f1-score   support  
  
      0          0.85      0.88      0.86      208  
      1          0.86      0.83      0.85      192  
  
accuracy                          0.85      400  
macro avg       0.86      0.85      0.85      400  
weighted avg     0.86      0.85      0.85      400  
  
0.855
```

Abbildung 4.4: Konfusions-Matrix Texterkennungs-Experiment

Die Konfusionsmatrix mit einer für ein Experiment annehmbare Fehlerquote und einer Accuracy von 0.855.

Visualisierung durch ELI5

Die Python Bibliothek “ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions”, o.D. unterstützt einige ML Bibliotheken und bietet die Möglichkeit Schlüsselwörter welche für eine Klassifikation relevant sind direkt in dem Ursprungstext darzustellen.

Weight	Feature
0.0189 ± 0.0492	bad
0.0130 ± 0.0385	worst
0.0076 ± 0.0256	boring
0.0072 ± 0.0228	supposed
0.0065 ± 0.0210	nothing
0.0064 ± 0.0239	stupid
0.0064 ± 0.0210	plot
0.0057 ± 0.0196	reason
0.0056 ± 0.0206	ridiculous
0.0054 ± 0.0193	waste
... 1490 more ...	

Eine Übersicht der Top Features zeigt die Funktion “show_weights()”.

```
1 eli5.show_weights(classifier, vec=  
vectorizer, top=10)
```

Allerdings verhält sich ELI5 bei einer binären Klassifizierung so dass nur eine Klasse (in diesem Fall ‘neg’) dargestellt wird. Die Farbe Grün stellt immer die aktuell gewählte Klasse dar weshalb hier auch negative Wörter grün eingefärbt sind.

Abbildung 4.5: Top Features Film Review Klassifizierung

Um einen Datensatz zu visualisieren verwendet man die Methode “explain_prediction()”, welche sowohl Details über die Features als auch eine Darstellung des Textes mit den hervorgehobenen Schlüsselwörtern anzeigt. Je nach verwendetem Modell weicht die Darstellung von dem hier dargestellten Bild ab, bei gewissen Tree Algorithmen (z.Bsp. DecisionTree) wird zusätzlich noch die Baumstruktur angezeigt.

```

1 doc = documents[414]
2 eli5.explain_prediction(classifier, doc, vec=vectorizer, target_names=['neg', 'pos'], top=20)

```

y=pos (probability 0.692) top features

Contribution?	Feature
+0.505	<BIAS>
+0.009	plot
+0.009	ni
+0.007	worst
... 884 more positive ...	
... 579 more negative ...	
-0.007	well
-0.036	Highlighted in text (sum)

b'susan granger's review of " the perfect storm " (warner bros .) \n " more people die on fishing boats , per capita , than working in any other job in the u . s . . \never journey a fishing boat makes can be an all-or-nothing risk . \nit is life at its most exhilarating and its most terrifying , " says director wolfgang petersen (" das boot ") . \nand that's just what he captures in this true story of struggle and humanity aboard a swordfishing boat , the andrea gail , sailing out of gloucester , massachusetts , in late october , 1991 . \nearly in bill wittliff's screenplay , based on sebastian junger's best-seller , we meet the crew of six . \nthe veteran captain (george clooney) is frustrated because he can't find fish on the grand banks , yet a rival skipper (mary elizabeth mastrantonio) brings in huge hauls . \nhis right-hand man (mark walhberg) needs money to build a new life with his girl-friend (diane lane) . \nthere's a devoted dad (john c. reilly) with an estranged wife and son , a free-spirited jamaican (allen payne) , a lonely guy (john hawkes) , and a last-minute replacement with a bad attitude (william fichtner) . \nthe skipper's convinced he can change his bad luck streak in remote flemish cap , and he does . \nbut then trouble begins . \nthere's a rogue wave , a man overboard and the ice machine breaks - with 60 , 000 lb . \nof fish that could spoil . \nbut that's minor compared with a deadly monster storm approaching which a boston meteorologist describes as " a disaster of epic proportions " that also threatens the lives of a coast guard helicopter rescue team trying to save three people stranded on a sailboat on the high seas . \nit's formulaic and there are cliches , but the walls of water , created by fluid dynamics simulating real-life phenomena , are awesome . \non the granger movie gauge of 1 to 10 , " the perfect storm " is a terrifying , suspenseful 8 . \nhang on for the white-knuckle thrill ride of the summer ! \n'

Abbildung 4.6: Visualisierung positives Film Review

Bei der Darstellung des negativen Film Reviews fällt auf dass Wörter welche für eine negative Stimmung stehen Grün markiert sind. Dies kommt daher dass für ELI5 die wahrscheinlichste Klasse 'neg' ist (81%) und deshalb alle Schlüsselwörter welche auf diese Klasse hinweisen grün markiert werden.

y=neg (probability 0.811) top features

Contribution?	Feature
+0.495	<BIAS>
+0.275	Highlighted in text (sum)
... 859 more positive ...	
... 583 more negative ...	
-0.007	worst
-0.007	ni
-0.008	plot
-0.014	bad

b'its a stupid little movie that trys to be clever and sophisticated , yet trys a bit too hard . \nwith the voices of woody allen , gene hackman , jennifer lopez , sylvester stallone , and sharon stone , this computer-animated yak-fest (think toy story [1996] filled with used merchandising) is one for the ant-eaters . \nthe main story is the independence of a worker named z (allen) . \nhe wants more to life than just digging away underground for the colony . \nwhen he finds out about ``insectopia , " a mythical place where all insects can run free , z , along with his colony's princess (stone) , journey out into the world to find a meaning for life . \nabout 15 minutes into the picture , i began to wonder what the point of the film was . \nhalfway through , i still didn't have an answer . \nby the end credits , i just gave up and ran out . \nnantz is a mindless mess of poor writing and even poorer voice-overs . \nallen is nonchalant , while i would have guessed , if i hadn't seen her in the mighty and basic instinct , stone can't act , even in a cartoon . \nthis film is one for the bugs : unfunny and extremely dull . \nhey , a bug's life may have a good time doing antz in . \n'

Abbildung 4.7: Visualisierung negatives Film Review

Blackbox Visualisierung mit ELI5

Während in dem letzten Beispiel ein White Box Model angewendet wurde, kann "ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions", o.D. auch Black Box Modelle analysieren. Dazu verwendet "ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions", o.D. eine LIME implementation und die Vorhersage zu erklären.

5 Schwächen von ML Modellen erkennen

5.1 Diskriminierung durch Bias

5.2 Adversarial Attacks

5.3 Data Poisoning

6 Weiterentwicklung von XAI

7 Anhang

7.1 Source Code

7.1.1 Entscheidungsbaum Visualisierung mit sklearn und Graphviz

```
1 from sklearn.datasets import load_iris
2 from sklearn import tree
3 from sklearn import datasets
4
5 X, y = load_iris(return_X_y=True)
6 clf = tree.DecisionTreeClassifier()
7 clf = clf.fit(X, y)
8
9 iris = datasets.load_iris()
10
11 dot_data = tree.export_graphviz(clf, out_file=None,
12                                 feature_names=iris.feature_names,
13                                 class_names=iris.target_names,
14                                 filled=True, rounded=True,
15                                 special_characters=True)
16
17 # print tree as text
18 from sklearn.tree import export_text
19 r = export_text(clf, feature_names=iris['feature_names'])
20 print(r)
21
22 # print tree as colored top-down tree
23 import graphviz
24 graph = graphviz.Source(dot_data)
25 graph
26
27 # plot decision surface
28 import numpy as np
29 import matplotlib.pyplot as plt
30 # Parameters
31 n_classes = 3
32 plot_colors = "ryb"
33 plot_step = 0.02
34
35 for pairidx, pair in enumerate([[0, 1], [0, 2], [0, 3],
36                                 [1, 2], [1, 3], [2, 3]]):
37     # We only take the two corresponding features
38     X = iris.data[:, pair]
39     y = iris.target
40
41     # Train
42     dTree = tree.DecisionTreeClassifier().fit(X, y)
43
44     # Plot the decision boundary
45     plt.subplot(2, 3, pairidx + 1)
```

```

47     x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
48     y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
49     xx, yy = np.meshgrid(np.arange(x_min, x_max, plot_step),
50                           np.arange(y_min, y_max, plot_step))
51     plt.tight_layout(h_pad=0.5, w_pad=0.5, pad=2.5)
52
53     Z = dTree.predict(np.c_[xx.ravel(), yy.ravel()])
54     Z = Z.reshape(xx.shape)
55     cs = plt.contourf(xx, yy, Z, cmap=plt.cm.RdYlBu)
56
57     plt.xlabel(iris.feature_names[pair[0]])
58     plt.ylabel(iris.feature_names[pair[1]])
59
60     # Plot the training points
61     for i, color in zip(range(n_classes), plot_colors):
62         idx = np.where(y == i)
63         plt.scatter(X[idx, 0], X[idx, 1], c=color, label=iris.target_names[i],
64                     cmap=plt.cm.RdYlBu, edgecolor='black', s=15)
65
66 plt.suptitle("Decision surface of a decision tree using paired features")
67 plt.legend(loc='lower right', borderpad=0, handletextpad=0)
68 plt.axis("tight")

```

Listing 7.1: Decision Tree Visualisierung

<https://scikit-learn.org/stable/modules/tree.html>

7.1.2 Bild-Klassifikation mit tf-explain

Das folgende Programm erzeugt mit den Bibliotheken Tensorflow (2.0) und tf-explain und en Algorithmen “Grad CAM” und “Integrated Gradients” Visualisierungen einer Bild-Klassifizierung.

```

1 import tensorflow as tf
2 from keras.applications.vgg16 import VGG16
3 from keras.preprocessing.image import load_img
4 from keras.preprocessing.image import img_to_array
5 from keras.applications.vgg16 import preprocess_input
6 from keras.applications.vgg16 import decode_predictions
7
8 model = tf.keras.applications.vgg16.VGG16(weights="imagenet", include_top=True
9      )
10
11 #print(model.summary())
12
13 imageOrig = load_img('D:/Master Thesis/dogs-vs-cats/test/DSC05797.JPG',
14                       target_size=(224, 224))
15 imageArr = img_to_array(imageOrig)  #output Numpy-array
16
17 imageReshaped = imageArr.reshape((1, imageArr.shape[0], imageArr.shape[1],
18                                 imageArr.shape[2]))
19
20 image = preprocess_input(imageReshaped)
21 predictions = model.predict(imageReshaped)
22
23 import numpy as np
24 top5predictions = np.argsort(predictions)[0,::-1][:5]
25
26 labels = decode_predictions(predictions)
27
28 for label in labels[0]:
29     print('%s (%.2f%)' % (label[1], label[2]*100))

```

```

28 from tf_explain.core.grad_cam import GradCAM
29 from mpl_toolkits.axes_grid1 import ImageGrid
30
31 def createImageGrid(imageOrig, predictions, labels, explainer, explainerArgs):
32     camImages = [imageOrig]
33     fig = plt.figure(figsize=(20., 20.))
34     grid = ImageGrid(fig, 111, # similar to subplot(111)
35                      nrows_ncols=(2, 3),
36                      axes_pad=0.5, # pad between axes in inch.
37                      )
38     for class_index in top5predictions:
39         camImages.append(explainer.explain(class_index=class_index, **
40                                             explainerArgs))
41
42     i = -1
43     for ax, im in zip(grid, camImages):
44         # Iterating over the grid returns the Axes.
45         ax.set_xticks([])
46         ax.set_yticks([])
47         label = labels[0][i]
48         if i >= 0:
49             ax.set_title('%s (%.2f%%)' % (label[1], label[2]*100), fontsize
50 =20)
51         ax.imshow(im)
52         i = i + 1
53
54
55 explainer = GradCAM()
56 createImageGrid(imageOrig, predictions, labels, explainer, {'model': model,
57   'layer_name': 'block5_conv3', 'validation_data': data})
58
59 from tf_explain.core.gradients_inputs import GradientsInputs
60 explainer = GradientsInputs()
61 createImageGrid(imageOrig, predictions, labels, explainer, {'model': model,
62   'validation_data': (np.array([imageArr]), None)})
63
64 from tf_explain.core.integrated_gradients import IntegratedGradients
65
66 explainer = IntegratedGradients()
67 createImageGrid(imageOrig, predictions, labels, explainer, {'model': model,
68   'validation_data': (np.array([imageArr]), None)})

```

Listing 7.2: Visualisiertes Neuronales Netz mit Tensorflow und tf-explain

<https://github.com/sicara/tf-explain>

7.1.3 Visualisierung einer Klassifikation mit lime

<https://github.com/marcotcr/lime/blob/master/doc/notebooks/Tutorial%20-%20Image%20Classification%20Keras.ipynb>

```

1 import lime
2 from lime import lime_image
3
4 explainer = lime_image.LimeImageExplainer()
5
6

```

```

7 explanation = explainer.explain_instance(np.vstack([imageArr]), model.predict,
8     top_labels=5, hide_color=0, num_samples=1000)
9
10
11 camImages = [imageOrig]
12 fig = plt.figure(figsize=(20., 20.))
13 grid = ImageGrid(fig, 111, # similar to subplot(111)
14     nrows_ncols=(2, 3),
15     axes_pad=0.5, # pad between axes in inch.
16 )
17 for class_index in range(0,5):
18     temp, mask = explanation.get_image_and_mask(explanation.top_labels[
19         class_index], positive_only=True, num_features=5, hide_rest=True)
20     camImages.append(mark_boundaries(temp / 2 + 0.5, mask))
21
22 i = -1
23 for ax, im in zip(grid, camImages):
24     # Iterating over the grid returns the Axes.
25     ax.set_xticks([])
26     ax.set_yticks([])
27     label = labels[0][i]
28     if i >= 0:
29         ax.set_title('%s (%.2f%%)' % (label[1], label[2]*100))
30     ax.imshow(im)
31     i = i + 1
32
33 plt.show()

```

Listing 7.3: Visualisiertes Neuronales Netz mit Tensorflow und lime

Akronyme

CNN Convolutional Neural Network. 10

DNN Deep Neural Network. 10

GAM

Generalized Additive Models. 6, 8, 9

GLM

Generalized Linear Models. 6, 8, 9

LFR Learned fair representations. 6

LIME

Local interpretable model-agnostic explanations. 13, 15, 27

LRP Layer-wise Relevance Propagation. 13

M-GBM

Monotonic gradient boosting. 6

ML Machine Learning. 1, 3, 5, 13, 26

PATE

Private aggregation of teacher ensembles. 6

SBRL

Scalable Bayesian rule list. 6

SLIM

Supersparse linear integer models. 6

SVCCA

Singular Vector Canonical Correlation Analysis. 17

TCAV

Testing with Concept Activation Vectors. 16

XAI Explainable artificial intelligence. 1

Glossar

Decision Tree

Entscheidungsbaum, Familie von ML Algorithmen. 9

Explainable artificial intelligence

deutsch erklärbare künstliche Intelligenz, Methodiken um Menschen die Vorhersagen durch Modelle des maschinellen Lernens zu erläutern. . 10, 25

Grad CAM

Gradient-weighted Class Activation Mapping, Technik welche für eine Entscheidung relevanten Bildinhalte optisch hervorhebt. 12

Machine Learning

deutsch Maschinelles lernen. Ein künstliches System lernt aus Beispielen und kann diese nach Beendigung der Lernphase verallgemeinern. . 25

Abbildungsverzeichnis

1.1 Entwicklung des Machine Learning als Zeitachse.	2
2.1 Ablauf einer erklärbaren Machine Learning Anwendung	3
3.1 Quelle: https://github.com/h2oai/mli-resources	8
3.2 Bedingungen lineare Regression	9
3.3 Auschluss-Bedingungen lineare Regression	10
3.4 Entscheidungsbaum visualisiert.	11
3.5 Entscheidungsbaum als Flächen dargestellt	11
3.6 Darstellung relevanter Bildinhalte durch LIME	15
3.7 Darstellung Vorgehensweise TCAV	17
3.8 Vergleich Verschiedener Klassen mit SVCCA	18
4.1 fiktives Logo	20
4.2 Original Testbild Katze	21
4.3 Testbild Meerschweinchen	21
4.4 Konfusions-Matrix Texterkennungs-Experiment	28
4.5 Top Features Film Review Klassifizierung	28
4.6 Visualisierung positives Film Review	29
4.7 Visualisierung negatives Film Review	29

Tabellenverzeichnis

Literaturverzeichnis Artikel

- Friedman, J. H. & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics 2008, Vol. 2, No. 3, 916-954*, arXiv <http://arxiv.org/abs/0811.1679v1>. <https://doi.org/10.1214/07-AOAS148>
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. & Sayres, R. (2017). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *ICML 2018*, arXiv <http://arxiv.org/abs/1711.11279v5>.
- Oh, S. J., Schiele, B. & Fritz, M. (2019). Towards Reverse-Engineering Black-Box Neural Networks, 121–144. https://doi.org/10.1007/978-3-030-28954-6_7

- Pang, B. & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.
- Raghu, M., Gilmer, J., Yosinski, J. & Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability, arXiv <http://arxiv.org/abs/1706.05806v2>.
- Ras, G., van Gerven, M. & Haselager, P. (2018). Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges, 19–36. https://doi.org/10.1007/978-3-319-98131-4_2
- Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why Should I Trust You?". <https://doi.org/10.1145/2939672.2939778>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, arXiv <http://arxiv.org/abs/1610.02391v4>. <https://doi.org/10.1007/s11263-019-01228-7>
- Simonyan, K. & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv <http://arxiv.org/abs/1409.1556v6>.
- Tensorflow. (2018). VGG16 Modell für Imagenet Klassifikationen. *github.com*. https://github.com/tensorflow/tensorflow/blob/r1.8/tensorflow/python/keras/_impl/keras/applications/vgg16.py

Literaturverzeichnis Bücher

Explainable and Interpretable Models in Computer Vision and Machine Learning. (2018, 1. September). Springer-Verlag GmbH. https://www.ebook.de/de/product/33610206/explainable_and_interpretable_models_in_computer_vision_and_machine_learning.html

Linkverzeichnis

- Datenethikkommission der Bundesregierung Bundesministerium des Innern, f. B. u. H. (2019). Gutachten der Datenethikkommission. https://www.bmjjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_DE.pdf?__blob=publicationFile&v=2
- ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions. (o.D.). <https://eli5.readthedocs.io/>
- ImageNet Challenge. (o.D.). <http://www.image-net.org/challenges/LSVRC/>
- Kim, B. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) [ICML 2018]. <https://github.com/tensorflow/tcav>
- Raghu, M. (2017). Interpreting Deep Neural Networks with SVCCA. <https://ai.googleblog.com/2017/11/interpreting-deep-neural-networks-with.html>
- scikit-learn Machine Learning in Python. (o.D.). <https://scikit-learn.org/stable/index.html>
- Tensorflow. (o.D.). <https://www.tensorflow.org/>
- with Python, T. C. & Scikit-Learn. (2018). Beispiel einer Textanalyse mit Scikit-Learn. <https://stackabuse.com/text-classification-with-python-and-scikit-learn/>

Listings

7.1 Decision Tree Visualisierung	32
7.2 Visualisiertes Neuronales Netz mit Tensorflow und tf-explain	33
7.3 Visualisiertes Neuronales Netz mit Tensorflow und lime	34