ETH zürich

Benchmarking and Survey of Explanation Methods for Black Box Models

Francesco Bodria[1], Fosca Giannotti[2], Riccardo Guidotti[3], Francesca Naretto[1], Dino Pedreschi[3], and Salvatore Rinzivillo[2]

[1] Scuola Normale Superiore, Pisa, Italy, {name.surname}@sns.it
[2] ISTI-CNR, Pisa, Italy, {name.surname}@isti.cnr.it
[3] Largo Bruno Pontecorvo, Pisa, Italy, {name.surname}@unipi.it

Abstract. The widespread adoption of black-box models in Artificial Intelligence has enhanced the need for explanation methods to reveal how these obscure models reach specific decisions. Retrieving explanations is fundamental to unveil possible biases and to resolve practical or ethical issues. Nowadays, the literature is full of methods with different explanations. We provide a categorization of explanation methods based on the type of explanation returned. We present the most recent and widely used explainers, and we show a visual comparison among explanations and a quantitative benchmarking.

Keywords: Explainable Artificial Intelligence, Interpretable Machine Learning, Transparent Models

1 Introduction

Today AI is one of the most important scientific and technological areas, with a tremendous socio-economic impact and a pervasive adoption in many fields of modern society. The impressive performance of AI systems in prediction, recommendation, and decision making support is generally reached by adopting complex Machine Learning (ML) models that "hide" the logic of their internal processes. As a consequence, such models are often referred to as "black-box models" [59,47,95]. Examples of black-box models used within current AI systems include deep learning models and ensemble such as bagging and boosting models. The high performance of such models in terms of accuracy has fostered the adoption of non-interpretable ML models even if the opaqueness of black-box models may hide potential issues inherited by training on biased or unfair data [77]. Thus there is a substantial risk that relying on opaque models may lead to adopting decisions that we do not fully understand or, even worse, violate ethical principles. Companies are increasingly embedding ML models in their AI products and applications, incurring a potential loss of safety and trust [32]. These risks are particularly relevant in high-stakes decision making scenarios, such as medicine, finance, automation. In 2018, the European Parliament introduced in the GDPR[4] a set of clauses for automated decision-making in terms of *a right of explanation* for all individuals to obtain "meaningful explanations of the logic involved" when automated decision making takes place. Also, in 2019, the High-Level Expert Group on AI presented the ethics guidelines for trustworthy AI[5]. Despite divergent opinions among legals regarding these clauses [53,121,35], everybody agrees that the need for the implementation of such a principle is urgent and that it is a huge open scientific challenge.

Paper Presentation

04.05.2022
Andri Simeon

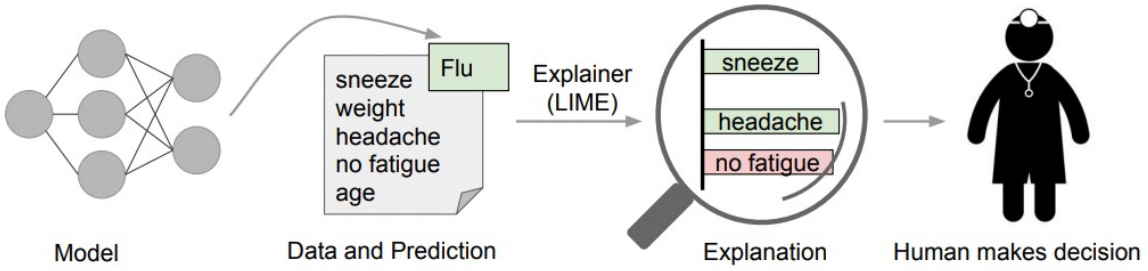
Agenda

1. Idea and Motivation of Explainers
2. Structure of Paper
3. Tabular Data
4. Image Data
5. Conclusions and Discussion

Idea and Motivation

eXplainable AI (XAI):

methods to produce or complement AI to make internal logic accessible and interpretable



Goals:

> developer: better models by identifying bias, errors and problems

> society: avoid violation of ethical principles

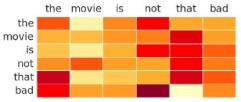
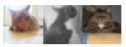
> stakeholder: reduce risks in e.g. medical and financial applications

Need:

> GDPR 2018: set of clauses by European Parliament -> right of explanation

Structure of Paper

Table 1: Examples of explanations divided for different data type and explanation

TABULAR	IMAGE	TEXT
Rule-Based (RB) A set of premises that the record must satisfy in order to meet the rule's consequence. $r = Education \leq College \rightarrow \leq 50k$	**Saliency Maps (SM)** A map which highlight the contribution of each pixel at the prediction. 	**Sentence Highlighting (SH)** A map which highlight the contribution of each word at the prediction. the movie is not that bad
Feature Importance (FI) A vector containing a value for each feature. Each value indicates the importance of the feature for the classification. 	**Concept Attribution (CA)** Compute attribution to a target "concept" given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)? 	**Attention Based (AB)** This type of explanation gives a matrix of scores which reveal how the word in the sentence are related to each other. 
Prototypes (PR) The user is provided with a series of examples that characterize a class of the black box $p = Age \in [35, 60]$, $Education \in [College, Master] \rightarrow$ "$\geq 50k$"	$p =$  $\rightarrow$ "cat"	$p =$ "... not bad ..." $\rightarrow$ "positive"
Counterfactuals (CF) The user is provided with a series of examples similar to the input query but with different class prediction $q = Education \leq College \rightarrow$ "$\leq 50k$" $c = Education \geq Master \rightarrow$ "$\geq 50k$"	$q =$  $\rightarrow$ "3" $c =$  $\rightarrow$ "8"	
$q =$ The movie is not that bad $\rightarrow$ "positive" $c =$ The movie is that bad $\rightarrow$ "negative"		

> Paper provides an extensive summary of explainer methods sorted by data structures

> Authors do not dive into architectures of methods but rather list them as survey

Tabular Data

Type	Name	Ref.	Authors	Year
FI	SHAP	[84]	Lundberg et al.	2007
	LIME	[102]	Ribeiro et al.	2016
	LRP	[17]	Bach et al.	2015
	DALEX	[19]	Biecek et al.	2020
	NAM	[6]	Agarwal et al.	2020
	CIU	[9]	Anjomshoae et al.	2020
	MAPLE	[99]	Plumb et al.	2018
RB	ANCHOR	[103]	Ribeiro et al.	2018
	LORE	[58]	Guidotti et al.	2018
	SLIPPER	[34]	Cohen et al.	1999
	LRI	[123]	Weiss et al.	2000
	MLRULE	[39]	Domingos et al.	2008
	RULEFIT	[48]	Friedman et al.	2008
	SCALABLE-BRL	[127]	Yang et al.	2017
	RULEMATRIX	[88]	Ming et al.	2018
	IDS	[78]	Lakkaraju et al.	2016
	TREPAN	[36]	Craven et al.	1996
	DECTEXT	[22]	Boz et al.	2002
	MSFT	[31]	Chipman et al.	1998
	CMM	[41]	Domingos et al.	1998
	STA	[132]	Zhou et al.	2016
	SKOPERULE	[48]	Gardin et al.	2020
	GLOCALX	[107]	Setzu et al.	2019
PR	MMD-CRITIC	[74]	Kim et al.	2016
	PROTODASH	[61]	Gurumoorthy et al.	2019
	TSP	[116]	Tan et al.	2020
	PS	[20]	Bien et al.	2011
CF	CEM	[40]	Dhurandhar et al.	2018
	DICE	[91]	Mothilal et al.	2020
	FACE	[100]	Poyiadzi et al.	2020
	CFX	[7]	Albini et al.	2020

Tabular Data

Datasets	Black-box Models	Explanation Types
German: Predicting credit risk of person based on attributes such as age, sex and job, among others Adult: Predicting salary of a person based on attributes such as age, work class and education, among others	Logistic Regression XGBoost CATBoost	**Feature Importance (FI)** **Rule-Based (RB)** Prototypes (PR) and Counterfactuals (CF) covered for images

Tabular Data: Feature Importance

> Popular explainers: SHAP, LIME

> Explainer assigns to each feature an importance value

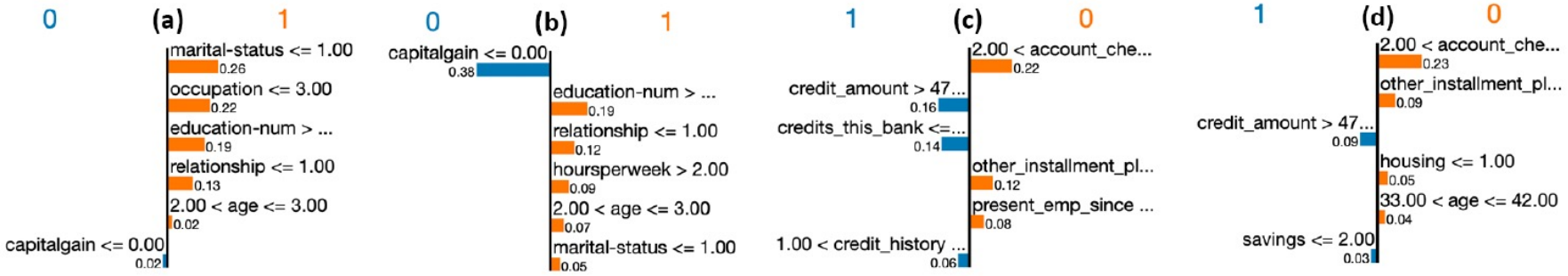
> Importance value represents how important particular feature was for prediction under analysis

> Example:

$$e = \{age = 0.8, income = 0.0, education = -0.2\}, y = deny$$

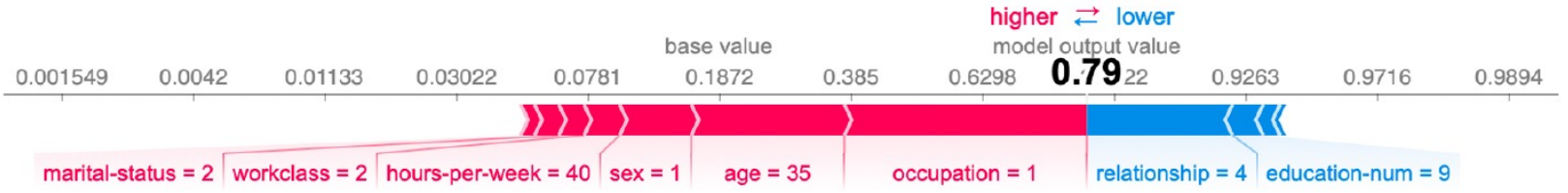
Feature Importance: LIME

adult: a) Logistic Regression, b) CATBoost **german**: c) Logistic Regression, d) XGB

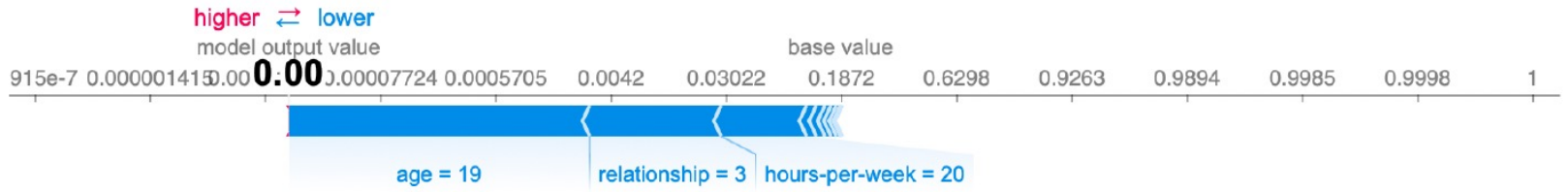


Tabular Data: SHAP Force Plot

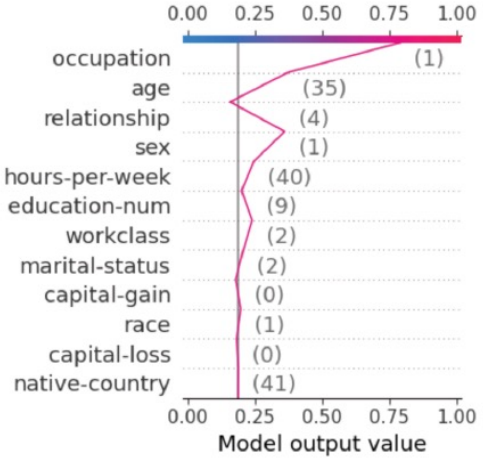
Class 1 (> 50k)

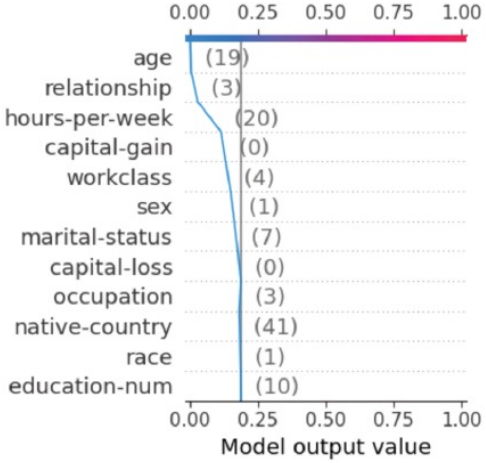


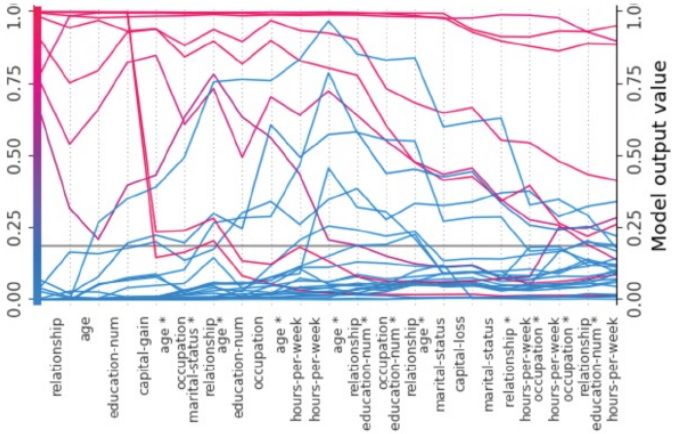
Class 0 (<= 50k)



Tabular Data: SHAP Decision Plot







Tabular Data: Rule-Based

> Popular explainers: ANCHOR, LORE

> represent if statements:

$r : p \rightarrow y, \qquad where\ p\text{: } premise, y\text{: } consequence, r\text{: } decision\ rule$

> ANCHOR: represents set of if statements, which fully describe rule no matter what other features look like

> LORE: consists of 1) factual decision rule and 2) set of counterfactual rules

> counterfactual rules need to be simple (short) and reasonable

$$x = \{\ Education = College,\ Occupation = Sales,\ Sex = Male,\ NativeCountry = US,\ Age = 19,\ Workclass = 2,\ HoursWeek = 15,\ Race = White,\ MaritialStatus = Married\text{-}civ,\ Relationship = Husband,\ CapitalGain = 2880,\ CapitalLoss = 0\ \},\ \leq 50k$$

$$r_{anchor} = \{Education \leq College,\ MaritialStatus > 1.00\ \} \rightarrow \leq 50k$$

$$r_{lore} = \{Education \leq Masters,\ Occupation > \text{-}0.34,\ HoursWeek \leq 40,\ WorkClass \leq 3.50,\ CapitalGain \leq 10000,\ Age \leq 34\} \rightarrow \leq 50k$$

$$c_{lore} = \{Education > Masters\ \} \rightarrow > 50k$$
$$\{CapitalGain > 20000\ \} \rightarrow > 50k$$
$$\{Occupation \leq \text{-}0.34\ \} \rightarrow > 50k$$

Quantitative Evaluation Measures

How good is our explanation method in approximating the black-box model?

Fidelity:

measures difference in accuracy between explanation and black-box method

Stability:

measures consistency of explanations on similar records in terms of Lipschitz Constant

$$L_x = max \frac{\|e_x - e_{x'}\|}{\|x - x'\|} \ \forall x' \in \mathrm{N}_x$$

$$\mathrm{N}_x : neighbourhood\ of\ x$$

Deletion/Insertion:

Insert/Remove features of the data in order of importance given by the explanation model and evaluate the modified data on black box model.

1) Faithfulness (deletion): compute correlation between feature importance and model performance

Tabular Data: Quantitative Comparison

Dataset	Black-Box	Fidelity				Faithfulness	
		LIME	SHAP	ANCHOR	LORE	LIME	SHAP
adult	LG	0.979	0.613	**0.989**	0.984	0.099 (0.30)	**0.38** (0.37)
	XGB	0.977	0.877	0.978	**0.982**	0.030 (0.32)	**0.36** (0.49)
	CAT	0.96	0.777	0.988	**0.989**	0.077 (0.32)	**0.44** (0.37)
german	LG	**0.984**	0.910	0.730	0.983	**0.23** (0.60)	0.19 (0.63)
	XGB	**0.999**	0.821	0.802	0.982	0.16 (0.26)	**0.44** (0.21)
	CAT	0.979	0.670	0.620	**0.981**	0.34 (0.33)	**0.43** (0.32)

Stability

Dataset	Black-Box	LIME	SHAP	ANCHOR	LORE
adult	LG	24.37 (2.74)	1.52 (4.49)	22.36 (8.37)	21.76 (11.80)
	XGB	10.16 (6.48)	2.17 (2.18)	26.53 (13.08)	30.01 (20.52)
	CAT	0.35 (0.43)	0.03 (0.01)	6.51 (4.40)	27.80 (70.05)
german	LG	18.87 (0.73)	19.01 (23.44)	101.07 (62.75)	622.12 (256.70)
	XGB	26.08 (14.50)	38.43 (30.66)	121.40 (98.43)	725.81 (337.26)
	CAT	2.49 (9.91)	15.92 (10.71)	123.79 (76.86)	756.70 (348.21)

Tabular Data: Qualitative Comparison

Feature Importance

Pro
> fast algorithm

Cons
> difficult to understand for non-experts

Rules, Counterfactuals

Pro
> human-readable
> counterfactuals tell you what to do instead

Cons
> slow algorithm

Image Data

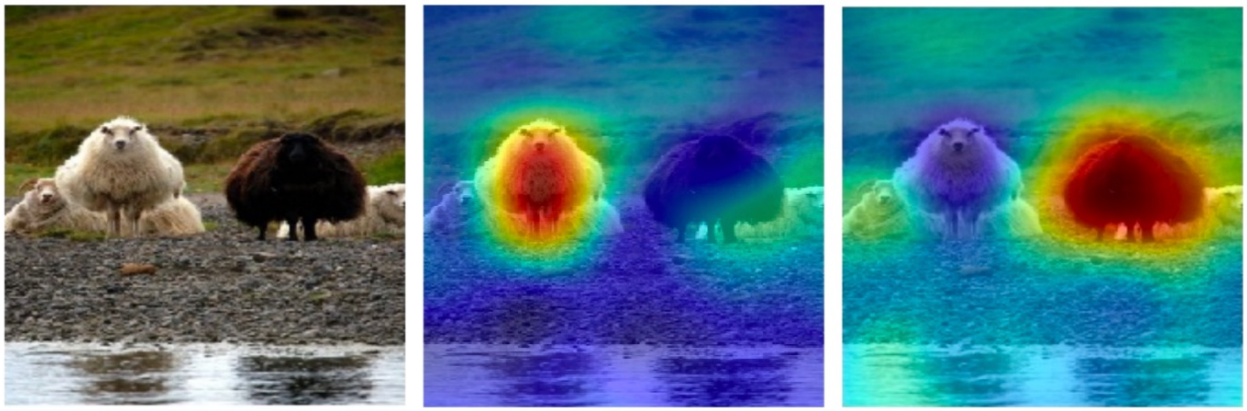
Type	Name	Ref.	Authors	Year
SM	SHAP	[84]	Lundberg et al.	2007
	LIME	[102]	Ribeiro et al.	2016
	ϵ-LRP	[17]	Bach et al.	2015
	INTGRAD	[115]	Sundararajan et al.	2017
	DEEPLIFT	[110]	Shrikumar et al.	2017
	SMOOTHGRAD	[112]	Smilkov et al.	2017
	XRAI	[70]	Kapishnikov et al.	2019
	GRADCAM	[106]	Selvaraju et al.	2017
	GRADCAM++	[27]	Chattopadhay et al.	2018
	RISE	[97]	Petsiuk et al.	2018
CA	TCAV	[75]	Kim et al.	2018
	ACE	[49]	Ghorbani et al.	2019
	CONCEPTSHAP	[129]	Yeh et al.	2020
	CACE	[54]	Goyal et al.	2019
CF	CEM	[40]	Dhurandhar, Amit, et al.	2018
	ABELE	[57]	Guidotti et al.	2020
	L2X	[29]	Chen et al.	2018
	GUIDED PROTO	[118]	Van Looveren et al.	2019
PR	MMD-CRITIC	[74]	Kim et al.	2016
	-	[76]	Koh et al.	2017
	PROTONET	[28]	Chen et al.	2019

Image Data

Datasets	Black-box Models	Explanation Types
MNIST: Handwritten digits dataset for classification CIFAR10: Images of 10 different classes to be classified Imagenet: Image database with more than 20’000 categories	CNN with two convolutions and two linear layers VGG16 (for imagenet)	**Saliency Maps (SM)** **Concept Attribution (CA)** **Prototypes (PR)** **Counterfactuals (CF)**

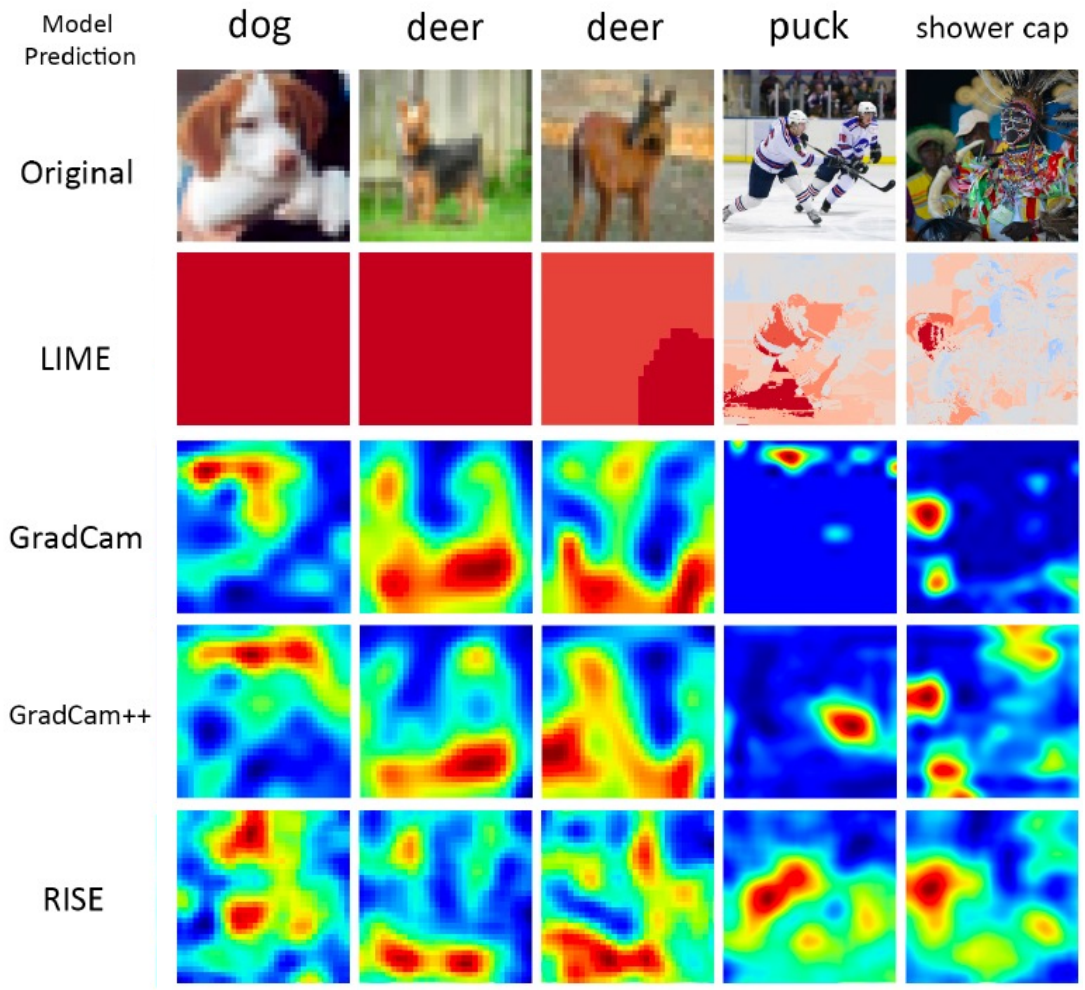
Image Data: Saliency Maps

> pixel‘s brightness represents how salient the pixel is



(a) Sheep - 26%, Cow - 17% (b) Importance map of ‘*sheep*’ (c) Importance map of ‘*cow*’

Saliency Maps: Qualitative Comparison



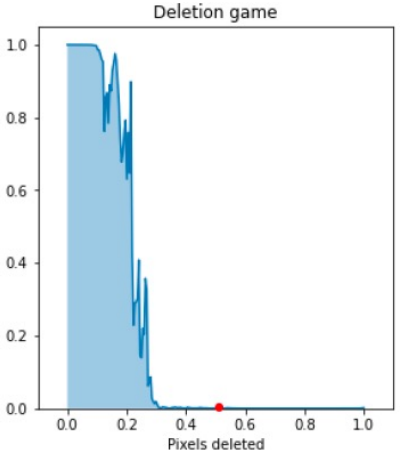
Saliency Maps: Quantitative Comparison

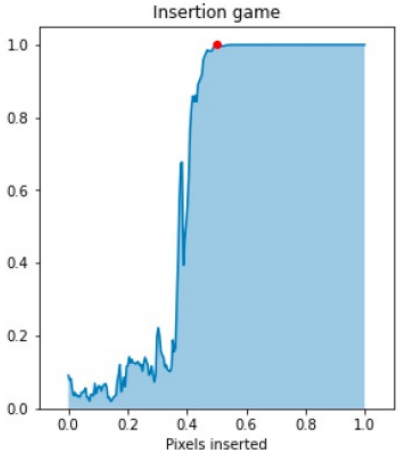
Deletion and insertion Metric:

> Substitute pixels in order of importance scores

> For every substitution query image to black-box model and obtain accuracy

> Metric: area under the curve of accuracy (AUC) as function of percentage of remove pixels





Saliency Maps: Quantitative Comparison

Insertion

	mnist	cifar	imagenet
LIME	0.807 (0.14)	0.41 (0.21)	0.34 (0.25)
ϵ-LRP	0.976 (0.02)	0.56 (0.20)	0.28 (0.19)
INTGRAD	0.973 (0.03)	0.50 (0.22)	0.27 (0.23)
DEEPLIFT	0.976 (0.02)	0.57 (0.20)	0.28 (0.19)
SMOOTHGRAD	**0.979** (0.03)	0.55 (0.23)	0.34 (0.26)
XRAI	0.956 (0.04)	0.58 (0.21)	0.40 (0.26)
GRADCAM	0.941 (0.04)	0.57 (0.20)	0.21 (0.19)
GRADCAM++	0.941 (0.04)	0.52 (0.22)	0.32 (0.26)
RISE	0.978 (0.03)	**0.61** (0.21)	**0.50** (0.26)

Deletion

	mnist	cifar	imagenet
LIME	0.388 (0.21)	0.221 (0.19)	0.051 (0.05)
ϵ-LRP	**0.120** (0.01)	0.127 (0.11)	**0.014** (0.02)
INTGRAD	0.126 (0.01)	0.148 (0.17)	0.029 (0.04)
DEEPLIFT	**0.120** (0.01)	0.127 (0.11)	**0.014** (0.02)
SMOOTHGRAD	0.135 (0.04)	0.153 (0.13)	0.033 (0.05)
XRAI	0.151 (0.04)	0.144 (0.07)	0.086 (0.11)
GRADCAM	0.297 (0.20)	0.153 (0.12)	0.139 (0.12)
GRADCAM++	0.252 (0.13)	0.283 (0.24)	0.081 (0.10)
RISE	**0.120** (0.01)	**0.124** (0.07)	0.044 (0.05)

Image Data: Concept Attribution

> explanation based on human-defined concepts rather than low-level features

> assigns a score to each concept based on the prediction

> example: quantifies how much the concept of stripes contributed to the class prediciton of zebra

image **concepts** **concept scores**

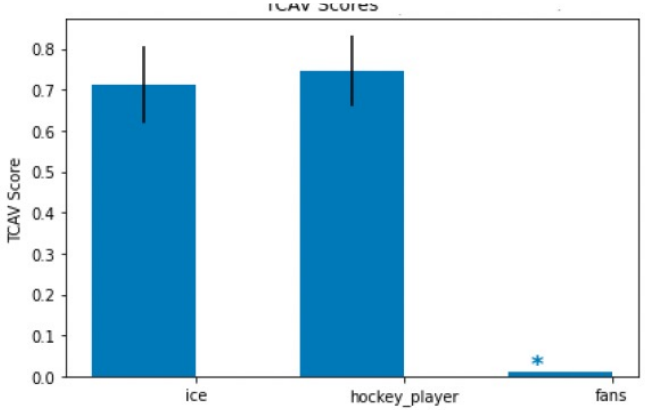


Image Data: Prototypes

> produce prototypical images that best represent a particular class, e.g cluster centroids

> example method: MMD-CRITIC analyses distribution of dataset under analysis, prototypes are instances near to dataset distribution whereas criticisms are fathest away

Criticism **Prototypes**

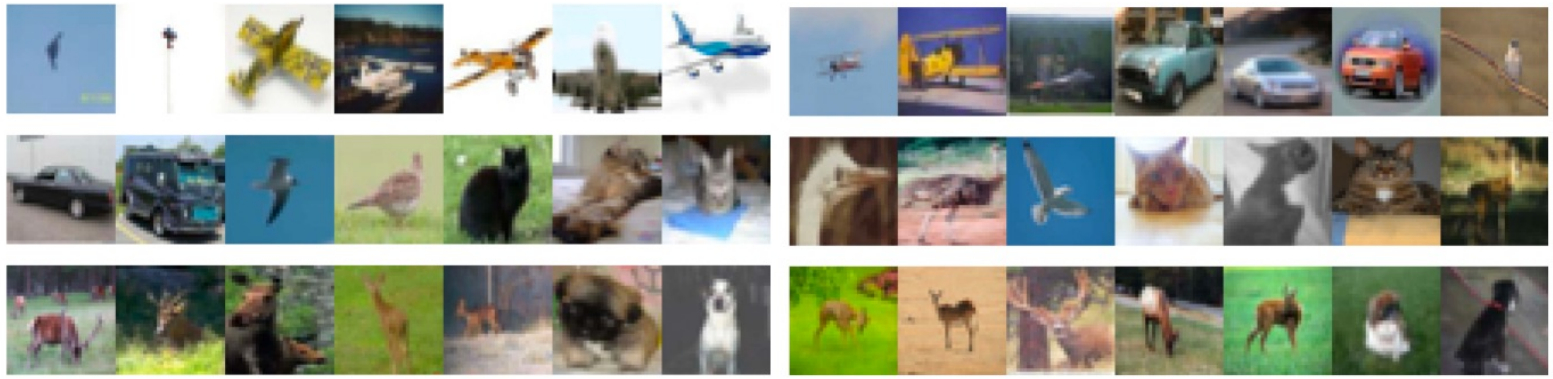
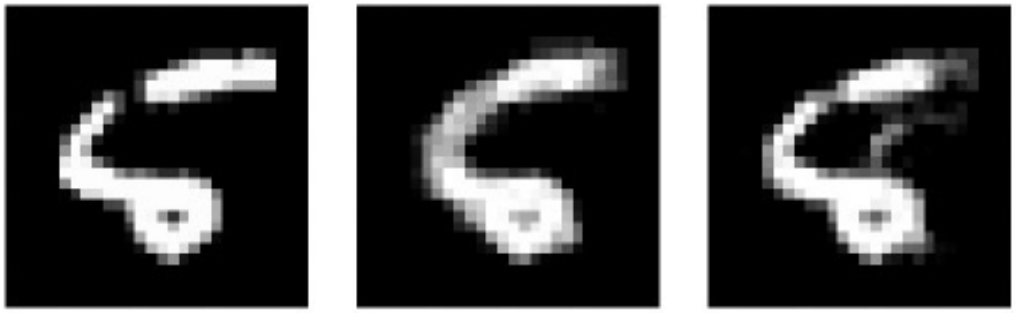


Image Data: Counterfactuals

> produce samples of images closed to the original one but with altered prediction



Query **CF 6** **CF 8**

Image Data: Qualitative Comparison

Saliency Maps:

low-level features -> great for debugging by human expert but not so for explaining

Concept Attribution:

high-level concepts -> human readable

recent type of explanation with potential improvements

Prototypes:

human reasoning often prototype based

Counterfactuals:

More user-friendly than prototypes since they highlight changes to make to obtain different prediction

Text Data

explanations at very early stage

compared to image and tabular data

Type	Name	Ref.	Authors	Year
SH	LIME	[102]	Ribeiro et al.	2016
	INTGRAD	[115]	Sundararajan et al.	2017
	L2X	[29]	Chen et al.	2018
	DEEPLIFT	[110]	Shrikumar et al.	2017
	LIONETS	[89]	Mollas et al.	2019
AB	-	[83]	Li et al.	2014
	EXBERT	[66]	Hoover et al.	2019
	-	[119]	Vaswani et al.	2017
Other	ANCHOR	[103]	Ribeiro et al.	2018
	QUINT	[2]	Abujabal et al.	2017
	CRIAGE	[98]	Pezeshkpour et al.	2019
	LASTS	[60]	Guidotti et al.	2020
	XSPELLS	[79]	Lampridis et al.	2020
	-	[101]	Rajani et al.	2019
	DOCTORXAI	[94]	Panigutti et al.	2020

Contributions

> look up table of explanation methods as starting point for research

> summarize what has been done so far in explainable AI in order to rise awareness for the need of more explanation and simplify/ push forward more research

Discussion

> it is very difficult to compare these models and to gain intuition whether they perform well

> there is a need for good and representative evaluation metrics

> gives a good overview over different approaches

> in terms layout, syntax and language the authors could have been more careful and clear, it seems as if they have weighted quantity over quality

Questions