Expert Systems and XAI

Professor Darian M. Onchis
https://staff.fmi.uvt.ro/~darian.onchis/
European Laboratory for Learning and Intelligent Systems (ELLIS)
darian.onchis@e-uvt.ro

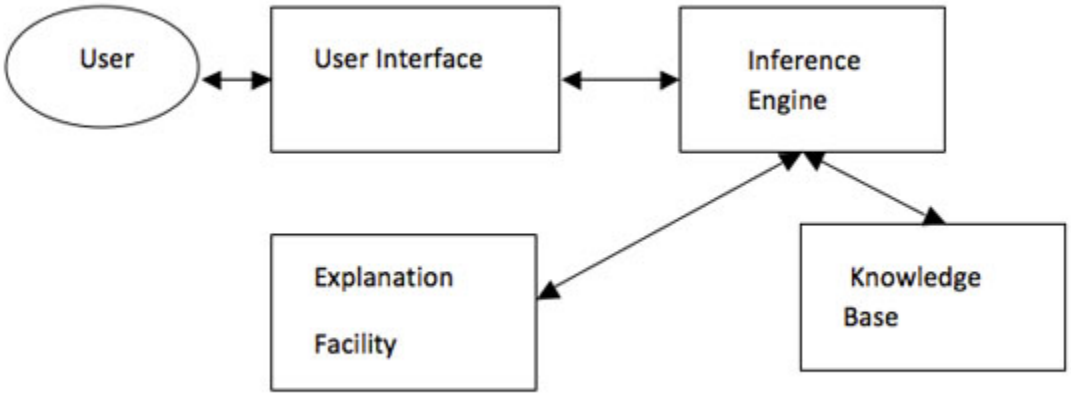
S·I·M·T = Signal, Image and Machine Learning Team

EXPLANATION FACILITIES USING SYMBOLIC AI

- **Symbolic AI systems store knowledge about their subject domains explicitly using symbols and manipulate these symbols using logical operatives.**
- Many expert systems used symbolic AI to manipulate knowledge stored in the form of rules. Explanations would then derive directly from these rules and provide a useful side-effect.

This explanation could be activated during or after a consultation with the system.

Fig. 1 provides a conceptual framework that shows the way explanation facilities work with symbolic expert system architectures. An end-user would communicate with the system via a user interface and an explanation facility which would interact with an expert system inference engine. The inference engine would use domain knowledge stored in the knowledge-base (often stored as rules) and control the consultation by determining what questions are to be asked in order to achieve its goals and derive conclusions or specify actions to be taken.



- **The explanation component would combine with the user-interface and knowledge-base to provide explanations. These explanations could take the form of the user wanting to know how the system advice was given – or why the system needs the answer to a question from a user.** For example, consider the following rule taken from a healthcare expert system:

- *RULE 1*
- *If alcoholic consumption is high*
- *And patient salt intake is high*
  - *Then blood pressure is likely to be high.*

This very simple example can be used to illustrate how explanation facilities work. If the above expert system arrived at the conclusion that the risk of heart failure is high, then the user could find out “how” the expert system arrived at that conclusion. The expert system might respond (all system responses shown in italic print) with something like the following:

- *I found patient alcohol consumption is high from user input*
  - *And I also found patient salt intake is high from user input*
  - *THEREFORE blood pressure is likely to be high from the activation of RULE 1.*

- A user could also find out why a particular question is being asked. For example, if the user is asked the question: *How many units of alcohol does the patient consume per week?*

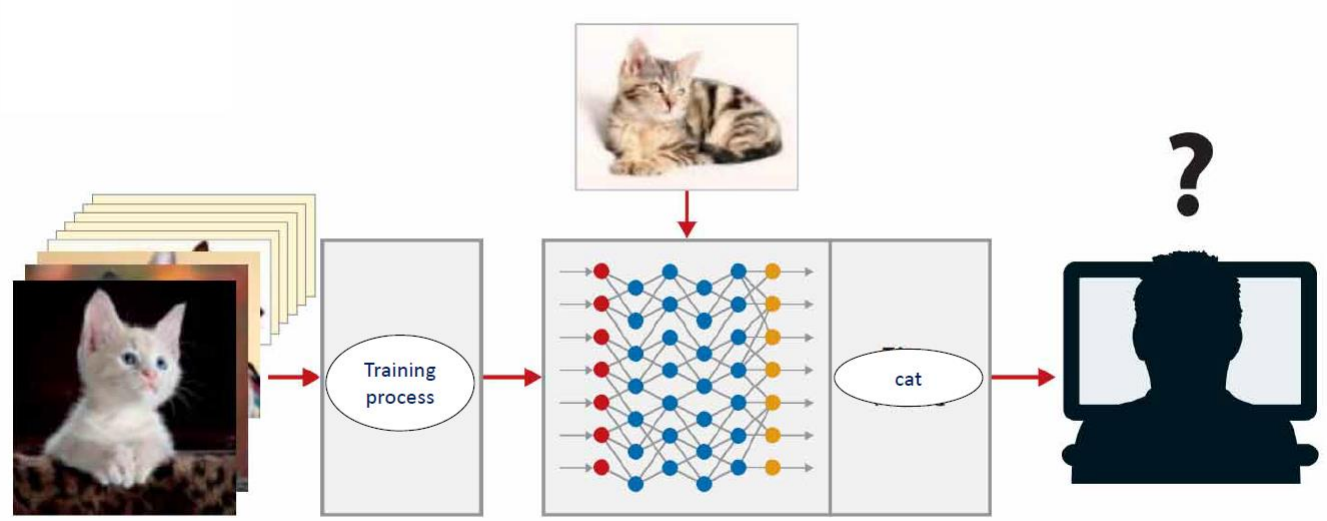
- The user could use explanation facilities to find out why it is asking this question. So user might respond: *Why?*

- The expert system might respond: *I am trying to prove RULE 1, to find out if blood pressure is likely to be high. To do this I need to find out if the alcohol consumption is high.*

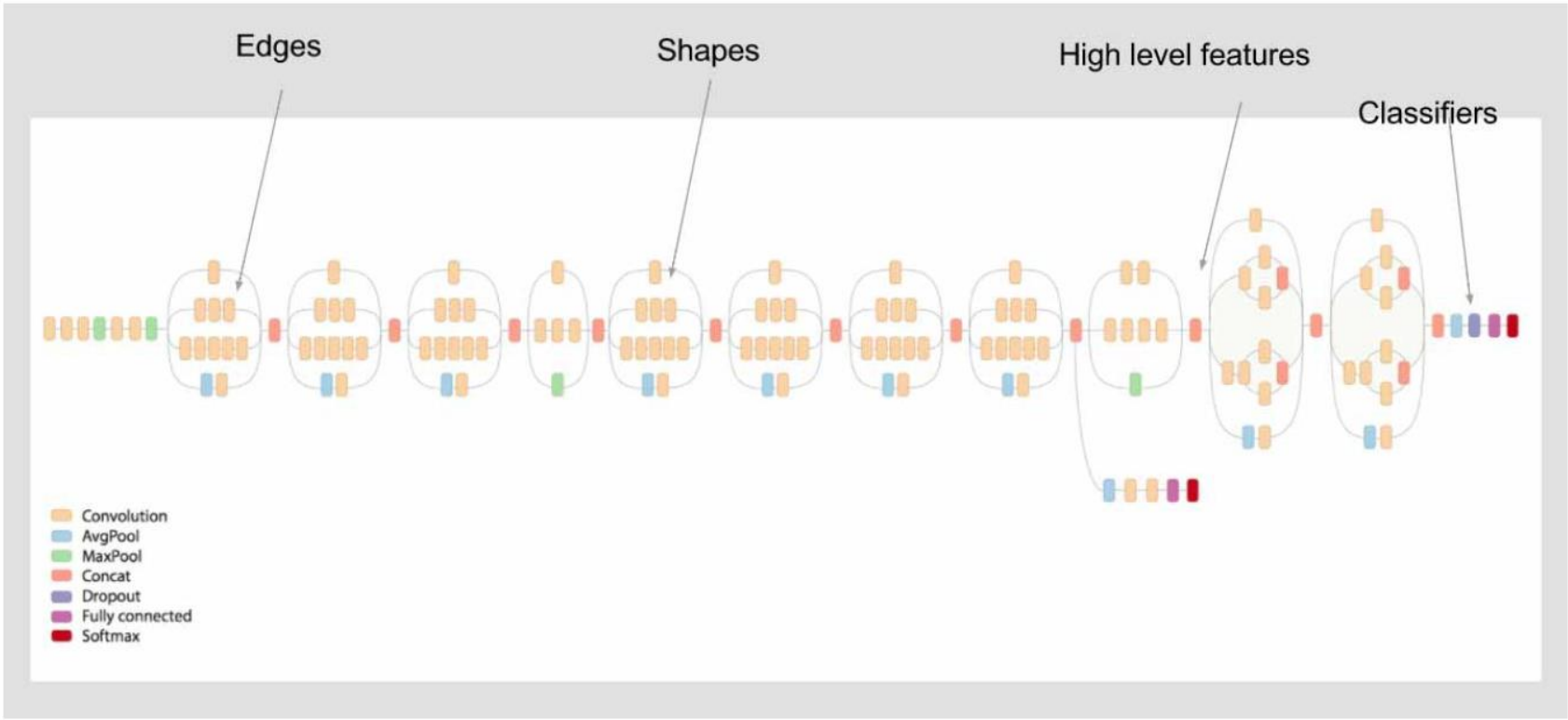
Back to nowadays AI

- Main machine learning paradigm predominantly uses neural networks –a non-symbolic AI paradigm that does not use explicit knowledge stored as rules of operation. Neural networks (NNs) work by learning from the use of large amounts of training data. Implicit knowledge is encoded in numeric parameters – called weights – and distributed all over the system.
- **This learning paradigm is not well suited to explanation because of the mathematical complexity of the network**. Nevertheless, a number of research models have been devised to incorporate explanation. Some use a decompositional approach for the extraction of rules from networks. This approach decomposes the network into single units, and then extract's rules to describe a unit's behavior.
- The Explainable AI project is more ambitious in that it seeks to integrate the problem-solving performance with explainabilty by using the same machine learning technology.
- **One approach to this could train the neural network to associate semantic attributes with hidden layer nodes – to enable learning of explainable features by modifying machine learning techniques**. For example, in learning to explain a neural network to identify birds from photographs, the semantic attributes would be things like "can fly" and "builds nests", and so on.

Why is cat ?

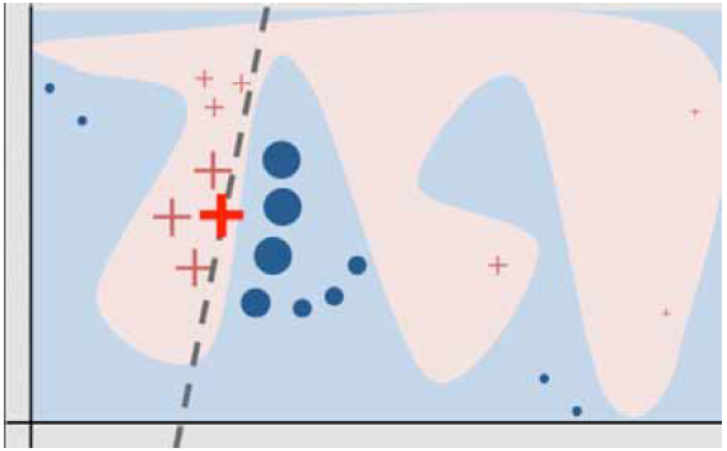


How CNN learn ?



What are explanations in the sense of LIME ?

- Explanation := local linear approximation of the model's behaviour. While the model may be very complex globally, it is easier to approximate it around the vicinity of a particular instance. While treating the model as a black box, we perturb the instance we want to explain and learn a sparse linear model around it -> used as explanation.



- Look at the image: The model's decision function is represented by the blue/pink background = clearly nonlinear. The bright red cross is the instance being explained (let's call it X). We sample instances around X, and weight them according to their proximity to X (weight here is indicated by size). We then learn a linear model (dashed line) that approximates the model well in the vicinity of X, but not necessarily globally!

https://github.com/marcotcr/lime

Computer Science > Machine Learning

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(Submitted on 16 Feb 2016 (v1), last revised 9 Aug 2016 (this version, v3))

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one. In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Machine Learning (stat.ML)

Cite as: arXiv:1602.04938 [cs.LG]

(or arXiv:1602.04938v3 [cs.LG] for this version)

Bibliographic data

Select data provider: **Semantic Scholar** | Prophy [Disable Bibex(What is Bibex?)]

References (19)

Data provided by: (report data issues) Semantic Scholar

Filter: Sort: Influence

Pages: 1 2 ▶ Skip: 1

Citations (1427)

Data provided by: (report data issues) Semantic Scholar

Filter: Sort: Influence

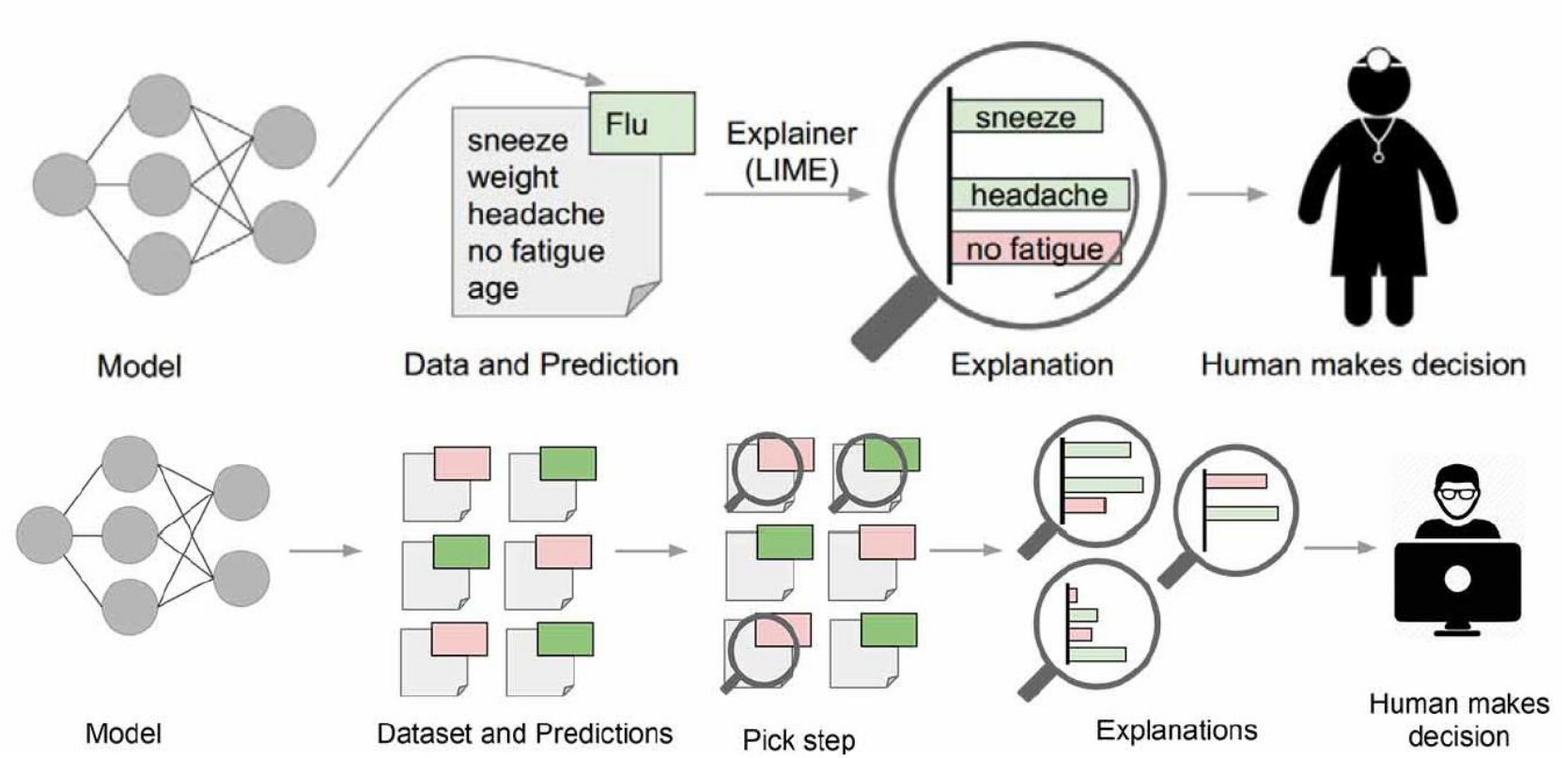
Pages: 1 2 3 4 5 ... 143 ▶ Skip: 1

Why **should** i **trust you**?: **Explaining** the **predictions** of **any classifier**

MT Ribeiro, S Singh, C Guestrin - Proceedings of the 22nd ACM ..., 2016 - dl.acm.org

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when ...

☆ Zitiert von: 2156 Ähnliche Artikel Alle 16 Versionen In EndNote importieren

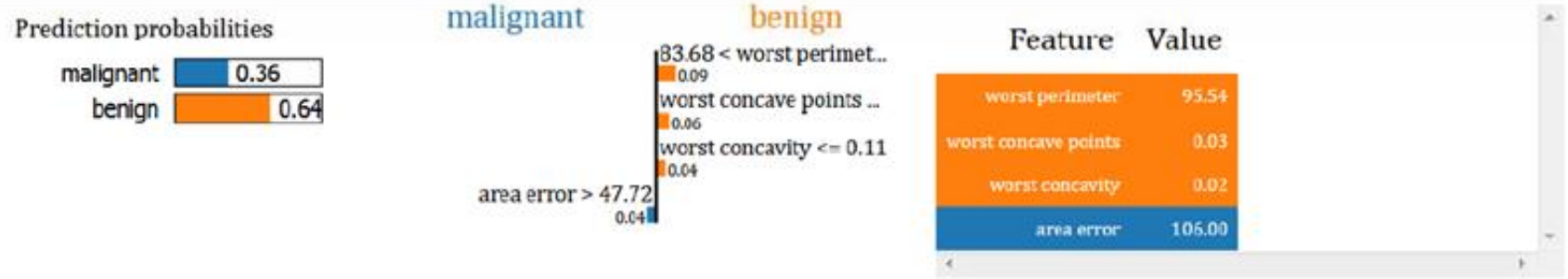


Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

In [12]:
```
explainer = lime.lime_tabular.LimeTabularExplainer(X_train, feature_names=breast.feature_names, class_names=breast.targe
```

Here we will take a sample from the test set (in this case the sample at index 76) and create an explainer instance for this sample. This will let us see why the algorithm made its prediction visually.

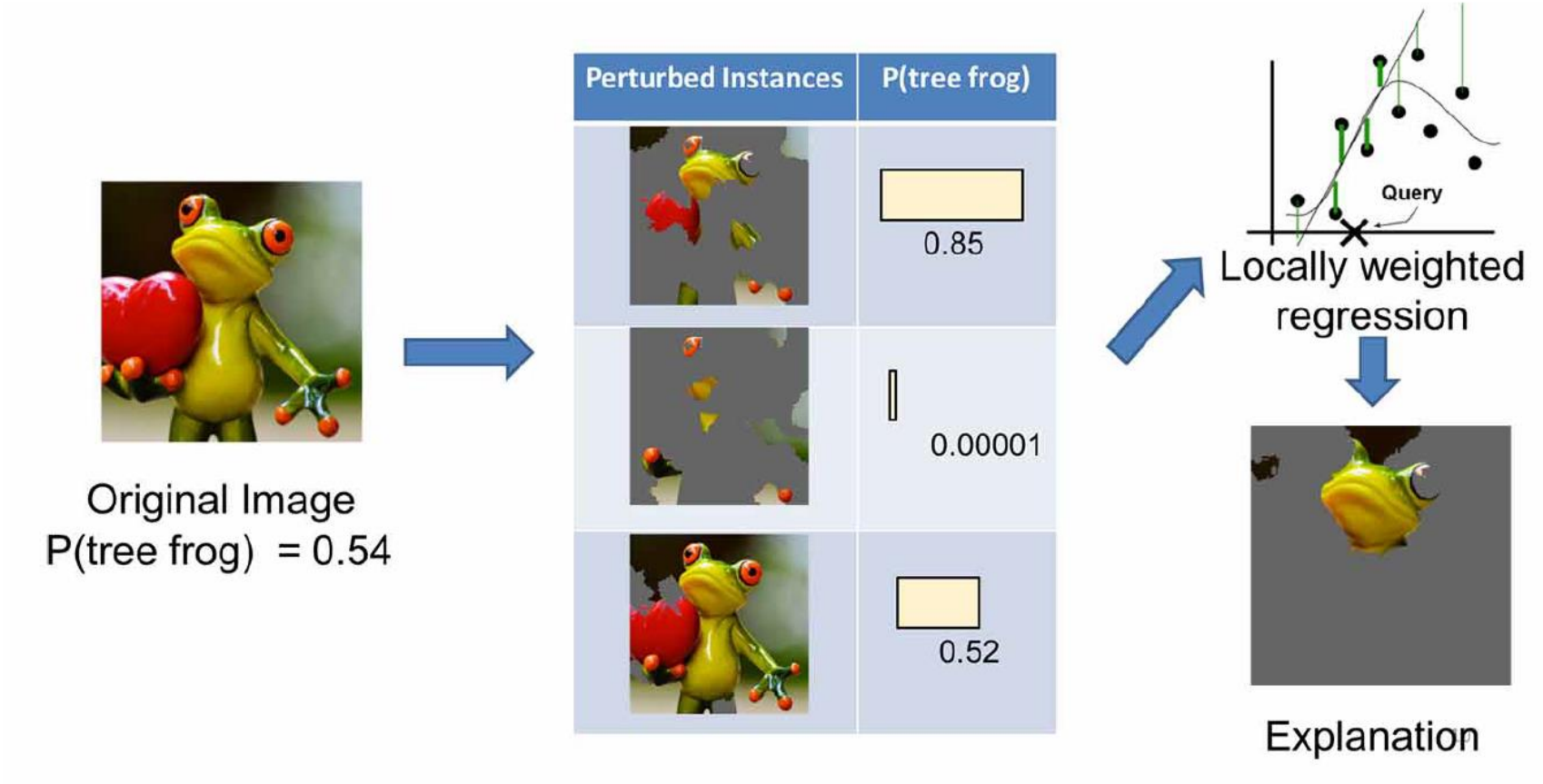
In [18]:
```
# For this demonstration, let's take the same sample each time, in this case sample index 86
i = 76
# For a random sample uncomment out the following line
# i = np.random.randint(0, X_test.shape[0])

exp = explainer.explain_instance(X_test[i], random_forest.predict_proba, num_features=4)
exp.show_in_notebook(show_table=True, show_all=False)
```



As you can see, the random forest algorithm has predicted with a probability of 0.64 that the sample at index 76 in the test set is benign

When using the explainer, we set the `num_features`parameter to 4, meaning the explainer shows the top 4 features that contributed to the prediction probabilities.

We chose 76 as it was a borderline decision. For example sample 86 is much more clear (this will we will set the `num_features` parameter to include all features so that we see each feature's contribution to the probability):

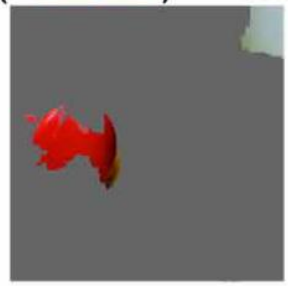


https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

P() = 0.54

P() = 0.07

P() = 0.05



https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME ALGORITHM

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), 2016 San Francisco (CA). ACM, 1135-1144, doi:10.1145/2939672.2939778.

LIME pros and cons

- + very popular,
- + many applications and contributors
- + model agnostic

- - local model behaviour can be unrealistic
- - unclear coverage
- - ambiguity (how to select the kernel width)

Black Box Explanations through Transparent Approximations, BETA

- BETA is a model agnostic approach to explain the behaviour of an (arbitrary) black box classifier (i.e. a function that maps a feature space to a set of classes) by simultaneously optimizing the accuracy of the original model and interpretability of the explanation for a human.
- Note: Interpretability and accuracy at the same time are difficult to achieve.
- Consequently, users are interactively integrated into the model and can thus explore the areas of black box models that interest them (most).

Computer Science > Artificial Intelligence

Interpretable & Explorable Approximations of Black Box Models

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, Jure Leskovec

(Submitted on 4 Jul 2017)

We propose Black Box Explanations through Transparent Approximations (BETA), a novel model agnostic framework for explaining the behavior of any black-box classifier by simultaneously optimizing for fidelity to the original model and interpretability of the explanation. To this end, we develop a novel objective function which allows us to learn (with optimality guarantees), a small number of compact decision sets each of which explains the behavior of the black box model in unambiguous, well-defined regions of feature space. Furthermore, our framework also is capable of accepting user input when generating these approximations, thus allowing users to interactively explore how the black-box model behaves in different subspaces that are of interest to the user. To the best of our knowledge, this is the first approach which can produce global explanations of the behavior of any given black box model through joint optimization of unambiguity, fidelity, and interpretability, while also allowing users to explore model behavior based on their preferences. Experimental evaluation with real-world datasets and user studies demonstrates that our approach can generate highly compact, easy-to-understand, yet accurate approximations of various kinds of predictive models compared to state-of-the-art baselines.

Comments: Presented as a poster at the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning

Subjects: **Artificial Intelligence (cs.AI)**

Cite as: arXiv:1707.01154 [cs.AI]

(or arXiv:1707.01154v1 [cs.AI] for this version)

Bibliographic data

References (9)

Citations (44)

Interpretable decision sets: A joint framework for description and prediction

H Lakkaraju, SH Bach, J Leskovec - Proceedings of the 22nd ACM ..., 2016 - dl.acm.org

One of the most important obstacles to deploying predictive models is the fact that humans do not understand and trust them. Knowing which variables are important in a model's prediction and how they are combined can be very powerful in helping people understand and trust automatic decision making systems. Here we propose interpretable decision sets, a framework for building predictive models that are highly accurate, yet also highly interpretable. Decision sets are sets of independent if-then rules. Because each rule can be ...

If Age <50 **and** Male =Yes:

If Past-Depression =Yes **and** Insomnia =No **and** Melancholy =No, **then** Healthy

If Past-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

If Age ≥ 50 **and** Male =No:

If Family-Depression =Yes **and** Insomnia =No **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

If Family-Depression =No **and** Insomnia =No **and** Melancholy =No **and** Tiredness =No, **then** Healthy

Default:

If Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression

If Past-Depression =No **and** Weight-Gain =Yes **and** Tiredness =Yes **and** Melancholy =Yes, **then** Depression

If Family-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

Algorithm 1 Optimization Procedure [5]

Input: Objective f, domain $\mathcal{ND} \times \mathcal{DL} \times C$, parameter δ, number of constraints k
$V_1 = \mathcal{ND} \times \mathcal{DL} \times C$
for $i \in \{1, 2 \cdots k+1\}$ **do** ▷ Approximation local search procedure
  $X = V_i$; $n = |X|$; $S_i = \emptyset$
  Let v be the element with the maximum value for f and set $S_i = v$
  while there exists a delete/update operation which increases the value of S_i by a factor of at least $(1 + \frac{\delta}{n^4})$ **do**
    Delete Operation: If $e \in S_i$ such that $f(S_i \backslash \{e\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \backslash e$

    Exchange Operation If $d \in X \backslash S_i$ and $e_j \in S_i$ (for $1 \leq j \leq k$) such that
    $(S_i \backslash e_j) \cup \{d\}$ (for $1 \leq j \leq k$) satisfies all the k constraints and
    $f(S_i \backslash \{e_1, e_2 \cdots e_k\} \cup \{d\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \backslash \{e_1, e_2, \cdots e_k\} \cup \{d\}$
  end while
  $V_{i+1} = V_i \backslash S_i$
end for
return the solution corresponding to $\max\{f(S_1), f(S_2), \cdots f(S_{k+1})\}$

Jon Lee, Vahab S Mirrokni, Viswanath Nagarajan & Maxim Sviridenko. Non-monotone submodular maximization under matroid and knapsack constraints. Proceedings of the forty-first annual ACM symposium on Theory of computing, 2009 (2018). ACM, 323-332.

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

Else if Risk-Depression=Yes **then** Depression

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

Else if Headaches=Yes **and** Dizziness=Yes, **then** Depression

Else if Doctor-Visits$\geq$ 0.3 **then** Diabetes

Else if Disposition-Tiredness=Yes **then** Depression

Else Diabetes

Notation	Definition	Term
$\mathcal{D}$	Input set of data points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$	Dataset
$\mathbf{x}$	Observed attribute values of a data point	
y	Class label of a data point	
$\mathcal{C}$	Set of class labels in $\mathcal{D}$	
p	(attribute, operator, value) tuple, *e.g.*, Age $\geq$ 50	Predicate
s	Conjunction of one or more predicates, *e.g.*, Age $\geq$ 50 **and** Gender = Female	Itemset
$\mathcal{S}$	Input set of itemsets	
r	Itemset-class pair (s, c)	Rule
$\mathcal{R}$	Set of rules $\{(s_1, c_1), \ldots, (s_k, c_k)\}$	Decision set

https://himalakkaraju.github.io

Himabindu Lakkaraju, Stephen H Bach & Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016. ACM, 1675-1684.

BETA, pros and cons

- + model agnostic
- + learns a compact two-level decision set
- + unambiguously

- - not so popular
- - unclear coverage
- - needs care

Layer-wise Relevance Propagation (LRP)

- LRP is general solution for understanding classification decisions by pixel-by-pixel (or layer-by-layer) decomposition of nonlinear classifiers.
- In a highly simplified way, LRP allows the "thinking processes" of neural networks to run backwards.
- Thereby it becomes comprehensible (for a human) which input had which influence on the respective result,
- e.g. in individual cases how the neural network came to a classification result, i.e. which input contributed most to the gained output.
- Example: If genetic data is entered into a network, it is not only possible to analyze the probability of a patient having a certain genetic disease, but with LRP also the characteristics of the decision.
- Such an approach is a step towards personalised medicine. In the future, such approaches will make it possible to provide an individual cancer therapy that is precisely "tailored" to the patient.

[HTML] On pixel-**wise** explanations for non-linear classifier decisions by **layer-wise relevance propagation**

S Bach, A Binder, G Montavon, F Klauschen... - PloS one, 2015 - journals.plos.org

Understanding and interpreting classification decisions of automated image classification systems is of high value in many applications, as it allows to verify the reasoning of the system and provides additional information to the human expert. Although machine learning ...

☆ 99 Zitiert von: 683 Ähnliche Artikel Alle 17 Versionen In EndNote importieren »

[PDF] iNNvestigate neural networks!

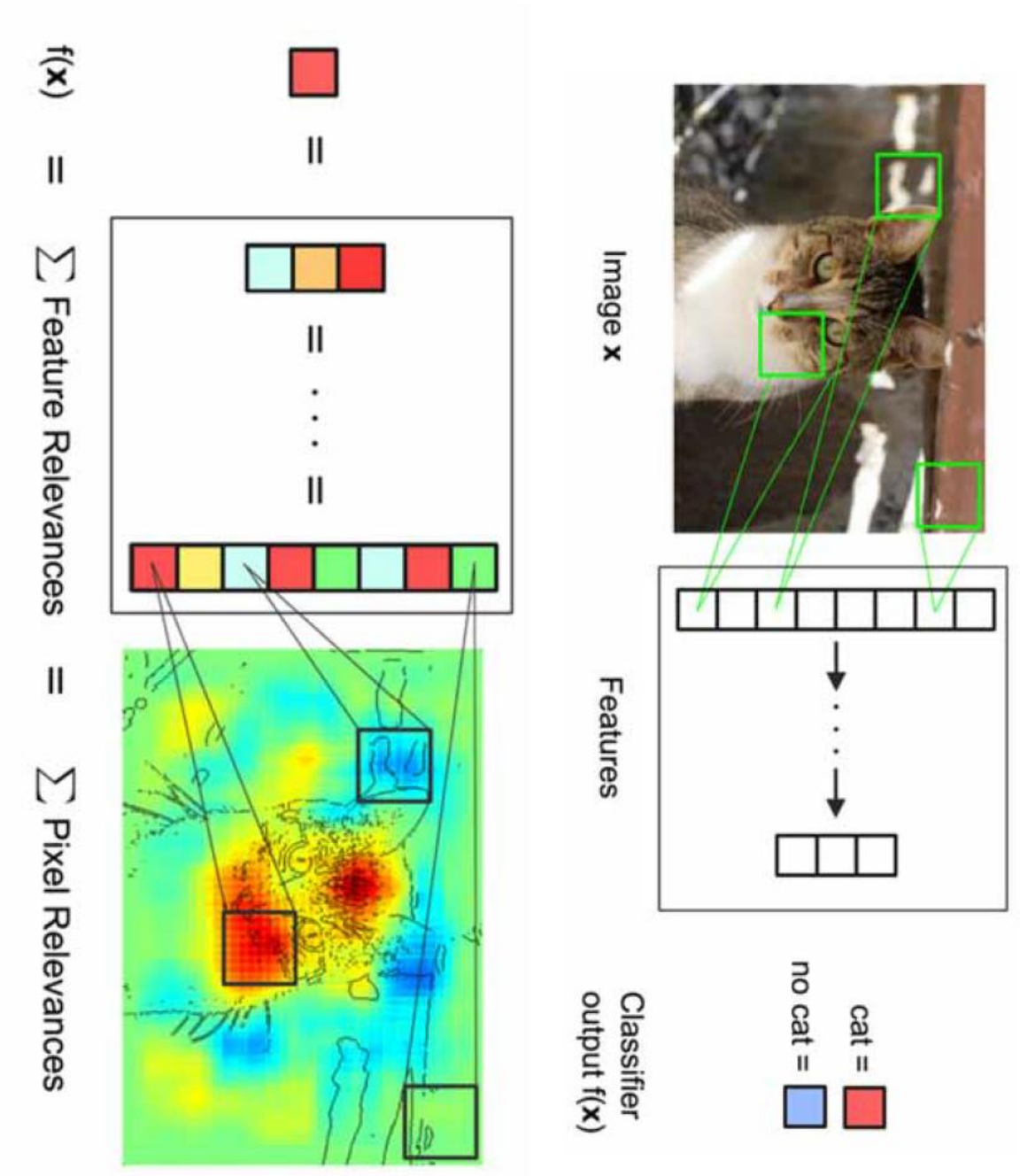
M Alber, S Lapuschkin, P Seegerer, M Hägele... - Journal of Machine ..., 2019 - jmlr.org

... On pixel-**wise** explanations for non-linear classifier decisions by **layer-wise relevance propagation**. PLOS ONE, 10(7):1–46, 2015 ... The **layer-wise relevance propagation** toolbox for artificial neural net- works. Journal of Machine Learning Research, 17:3938–3942, 2016b ...

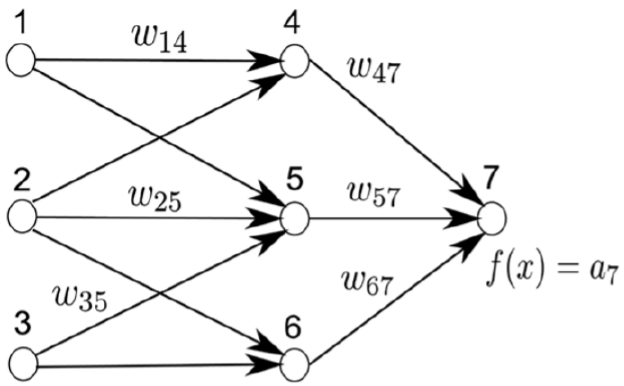
☆ 99 Zitiert von: 26 Ähnliche Artikel Alle 8 Versionen Web of Science: 1 In EndNote importieren

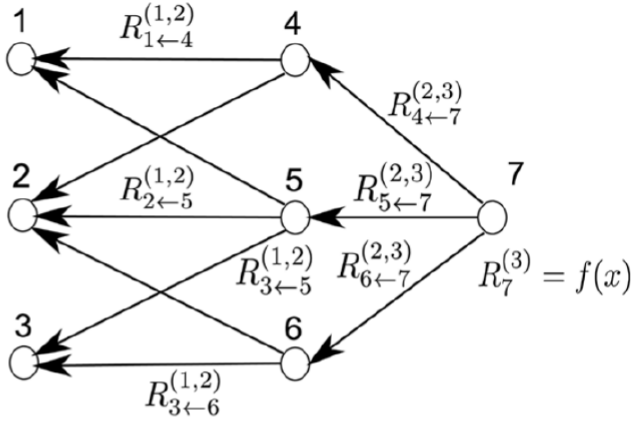
Grégoire Montavon 2019. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison. *In:* Samek, Wojciech, Montavon, Grégoire, Vedaldi, Andrea, Hansen, Lars Kai & Müller, Klaus-Robert (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Cham: Springer International Publishing, pp. 253-265, doi:10.1007/978-3-030-28954-6_13.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



NN classifier

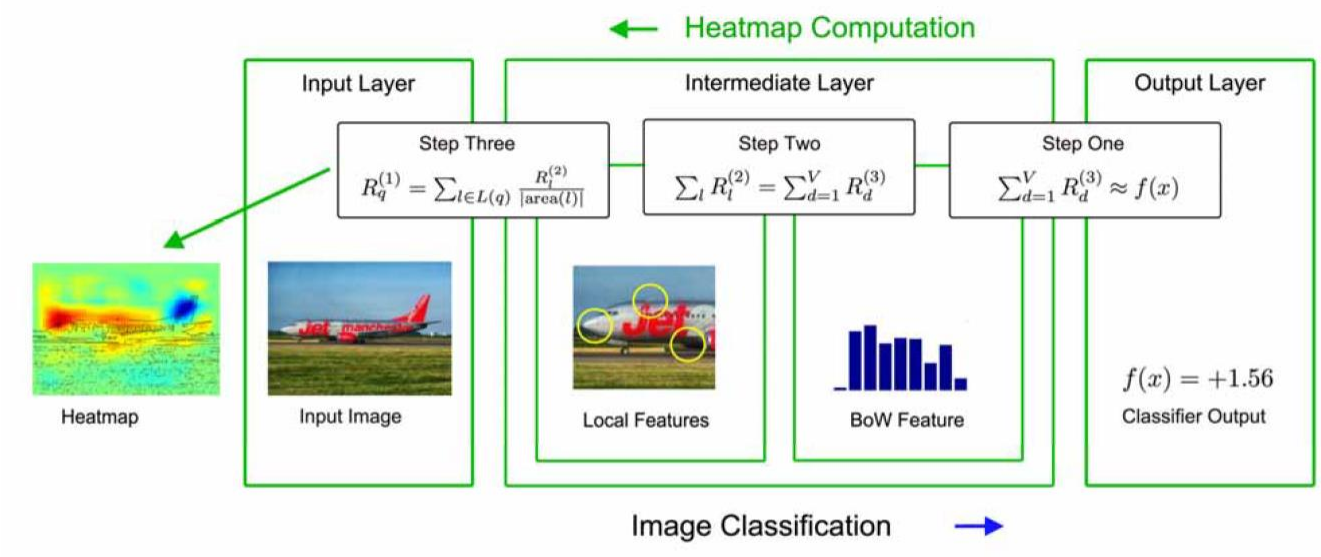




$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_d R_d^{(1)}$$

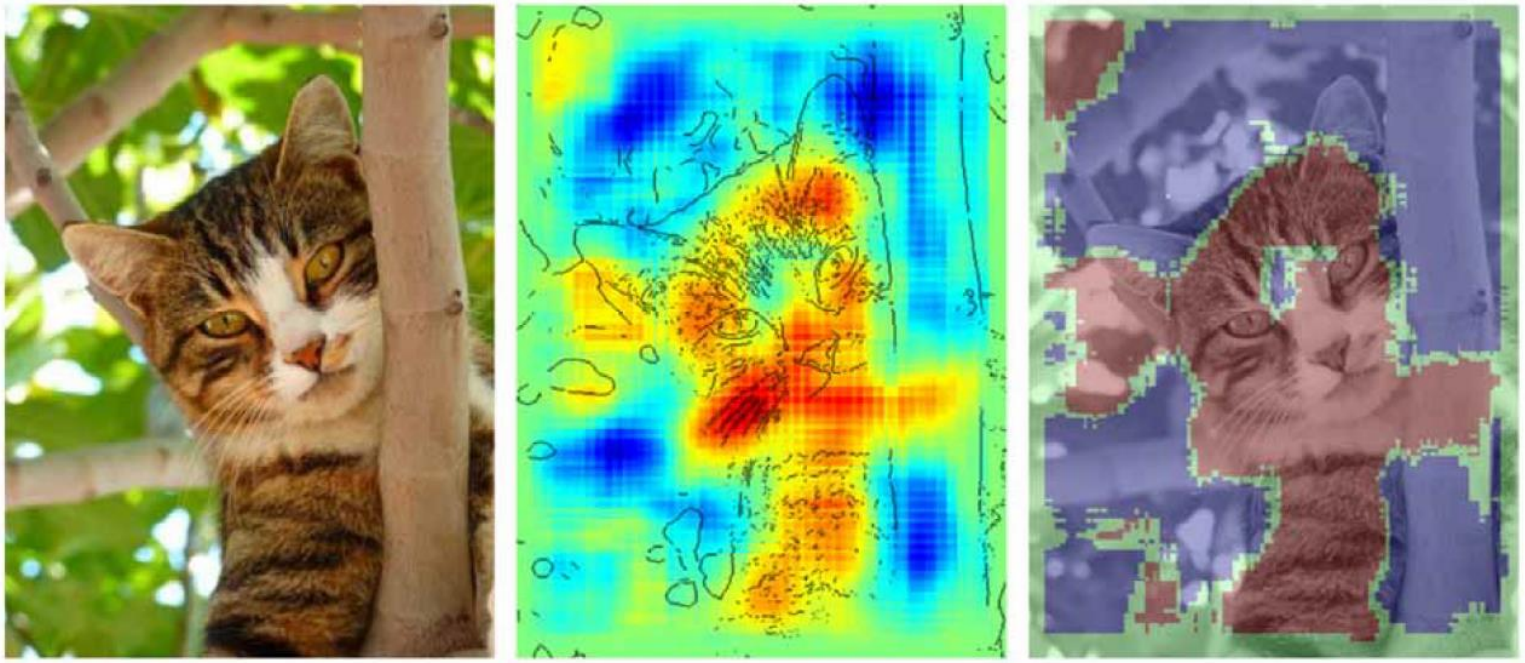
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

HEATMAP computation



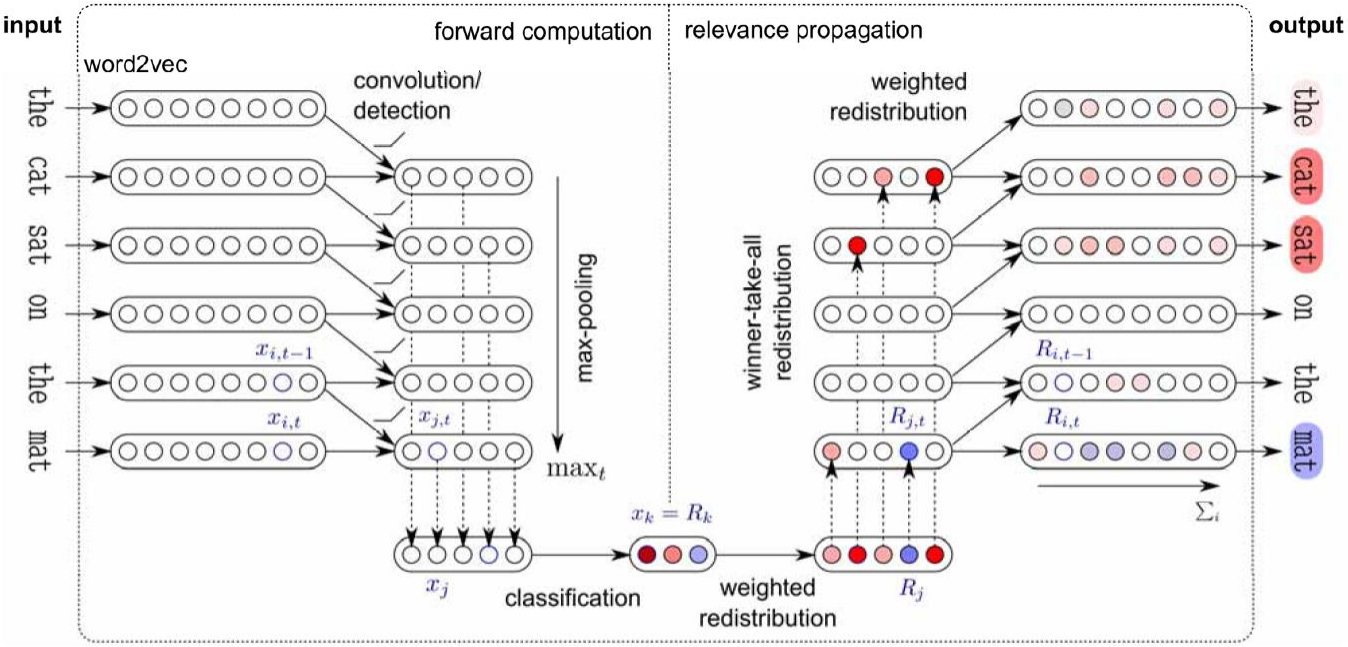
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

Pixel wise decomposition



Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

What is relevant in a text



Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller & Wojciech Samek 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one,* 12, (8), e0181142, doi:10.1371/journal.pone.0181142.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller & Wojciech Samek 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one,* 12, (8), e0181142, doi:10.1371/journal.pone.0181142.

Computer Science > Machine Learning

iNNvestigate neural networks!

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, Pieter-Jan Kindermans

(Submitted on 13 Aug 2018)

In recent years, deep neural networks have revolutionized many application domains of machine learning and are key components of many critical decision or predictive processes. Therefore, it is crucial that domain specialists can understand and analyze actions and pre- dictions, even of the most complex neural network architectures. Despite these arguments neural networks are often treated as black boxes. In the attempt to alleviate this short- coming many analysis methods were proposed, yet the lack of reference implementations often makes a systematic comparison between the methods a major effort. The presented library iNNvestigate addresses this by providing a common interface and out-of-the- box implementation for many analysis methods, including the reference implementation for PatternNet and PatternAttribution as well as for LRP-methods. To demonstrate the versatility of iNNvestigate, we provide an analysis of image classifications for variety of state-of-the-art neural network architectures.

Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as: arXiv:1808.04260 [cs.LG]
(or arXiv:1808.04260v1 [cs.LG] for this version)

Bibliographic data

Select data provider: **Semantic Scholar** | Prophy [Disable Bibex(What is Bibex?)]

References (28) Citations (20)

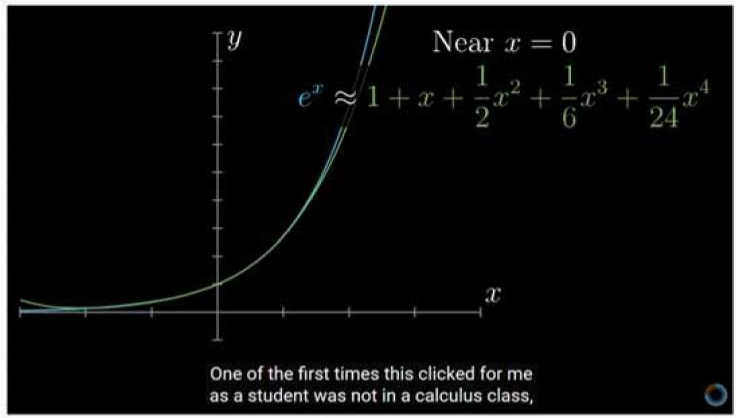
https://github.com/albermax/innvestigate

https://github.com/sebastian-lapuschkin/lrp_toolbox

https://github.com/ArrasL/LRP_for_LSTM

Also Explore:
https://innvestigate.readthedocs.io/en/latest/modules/analyzer.html#module-innvestigate.analyzer.relevance_based.relevance_analyzer

Deep Taylor Series



https://www.youtube.com/watch?v=3d6DsjIBzJ4

$$f(\boldsymbol{x}) = f(\tilde{\boldsymbol{x}}) + \left(\left.\frac{\partial f}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}}\right)^{\top} \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}}) + \varepsilon = 0 + \sum_p \underbrace{\left.\frac{\partial f}{\partial x_p}\right|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}} \cdot (x_p - \tilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon,$$

Brook Taylor

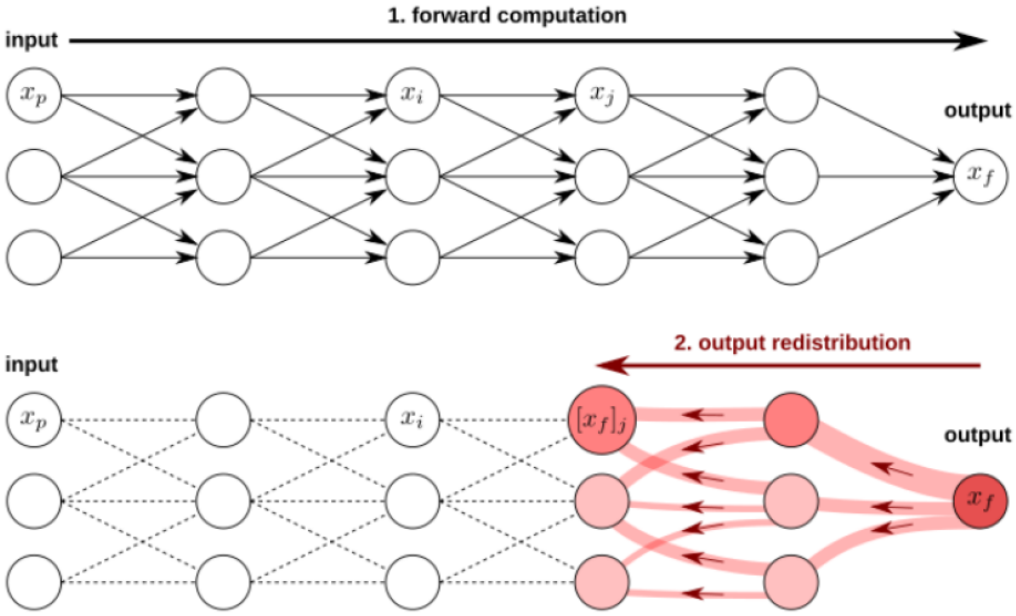
Brook Taylor (1685-1731)

Born	18 August 1685 Edmonton, Middlesex, England
Died	29 December 1731 (aged 46) London, England
Residence	England
Nationality	English
Alma mater	St John's College, Cambridge
Known for	Taylor's theorem Taylor series

https://en.wikipedia.org/wiki/Brook_Taylor

How it works ?

- running a backward pass on the NN using a predefined set of rules; produces decomposition of the NN output on the input variables.
- (1) dissociating the overall computation into a set of localized neuron computations, and
- (2) recombining these local computations



http://www.heatmapping.org/deeptaylor

Definition 1. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

$$\forall\, \boldsymbol{x}\colon f(\boldsymbol{x}) = \sum_{p} R_p(\boldsymbol{x}).$$

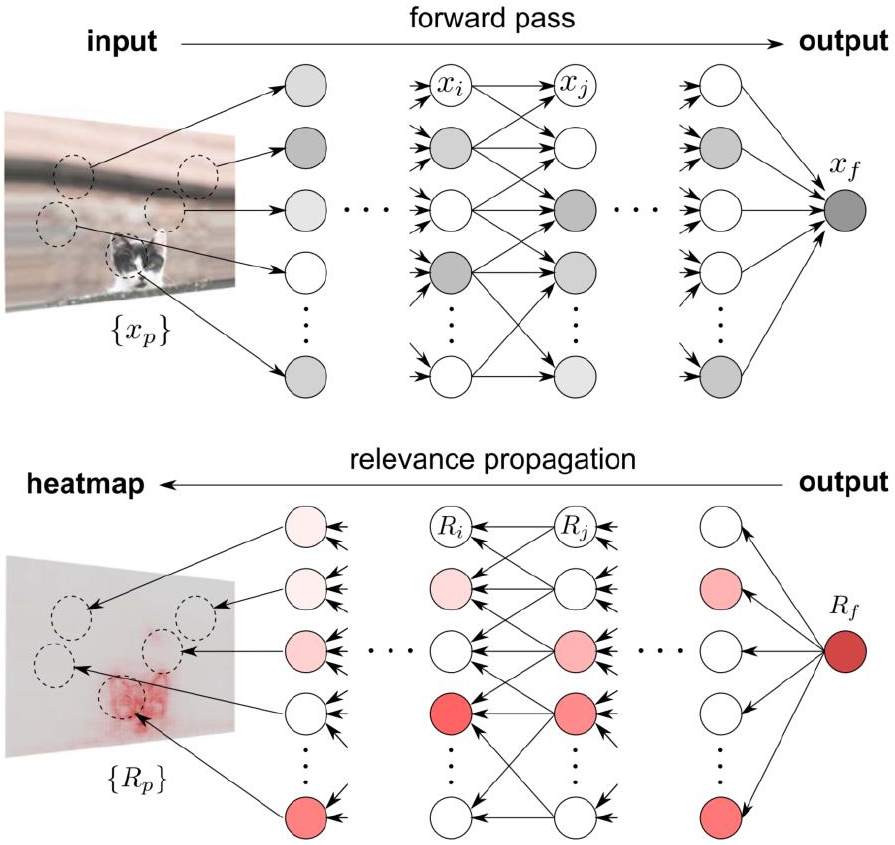
Definition 2. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *positive* if all values forming the heatmap are greater or equal to zero, that is:

$$\forall\, \boldsymbol{x}, p\colon R_p(\boldsymbol{x}) \geq 0$$

Definition 3. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *consistent* if it is conservative *and* positive. That is, it is consistent if it complies with Definitions 1 and 2.

Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.

Heatmaps again…



Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.

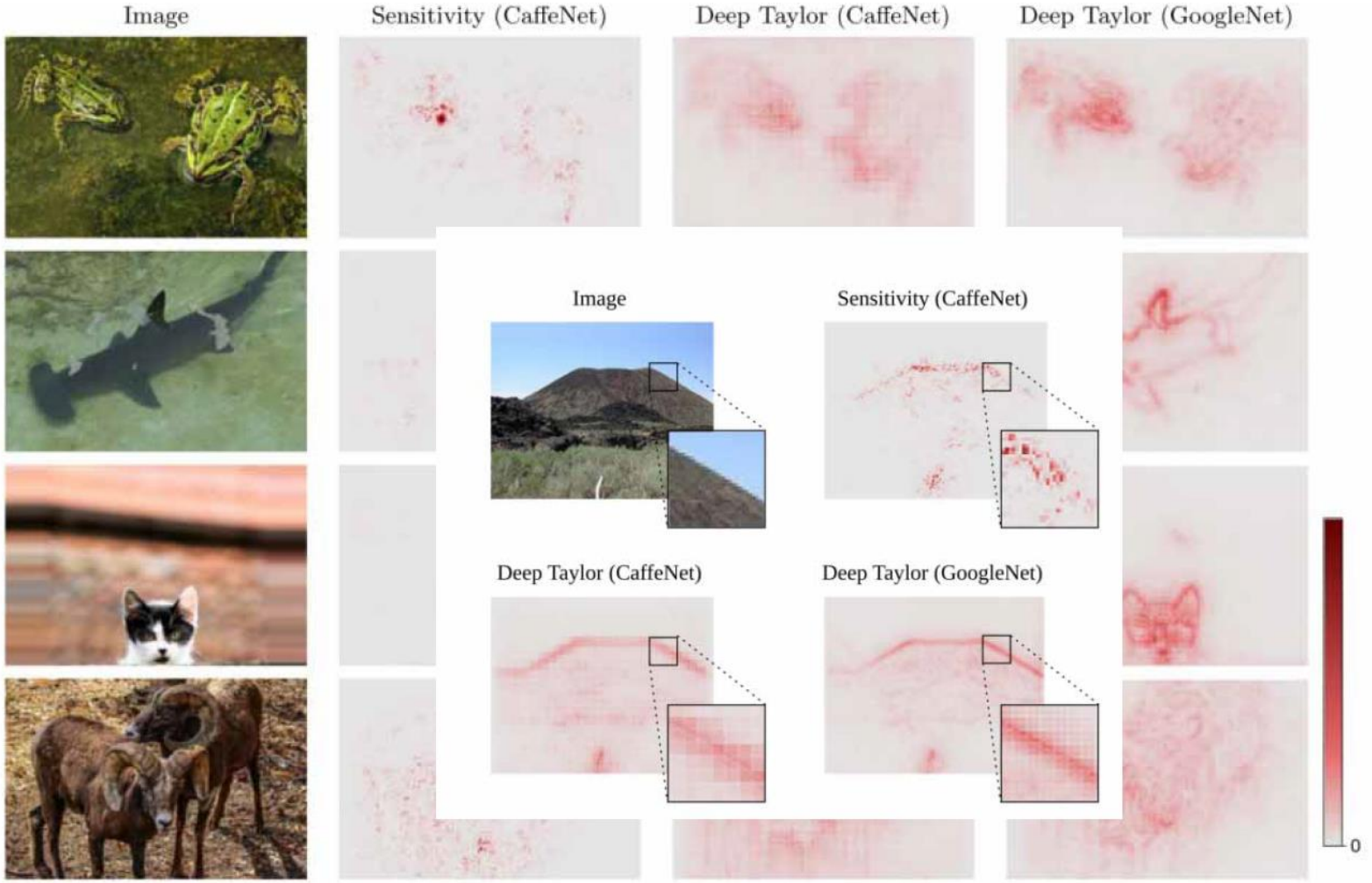
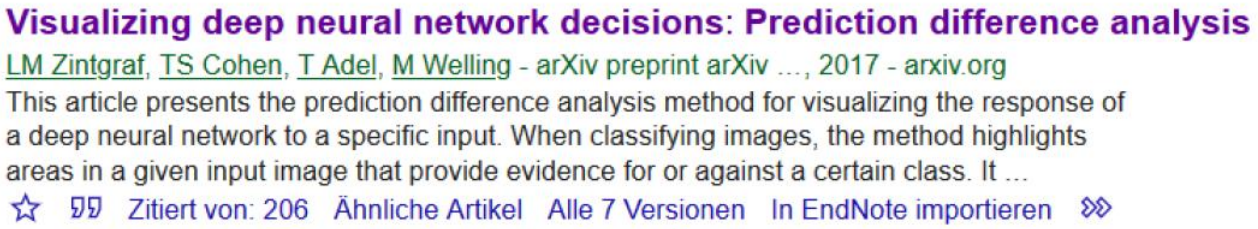


Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
0

Other methods

- **SHAP (SHapley Additive exPlanations)** is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (see [papers](https://github.com/slundberg/shap#citations) for details and citations.



Also IBM https://github.com/IBM/AIX360

Journals and conference

- https://www.journals.elsevier.com/expert-systems-with-applications
- https://ecmlpkdd2020.net/

- Self-Organising Maps: In Depth (for the colors project)
- DEEP LEARNING ALGORITHMS, https://theaisummer.com/Deep-Learning-Algorithms/?fbclid=IwAR1QOVfFSe8W4W-HrVQStZP2oaZtnwHe_sVEbvtpkmRSfm4KKhPTdkan0eg