Seminar Explainable AI
Module 5
Selected Methods Part 1
LIME-BETA-LRP-DTD-PDA

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)

Institute for Medical Informatics/Statistics, Medical University Graz, Austria

and

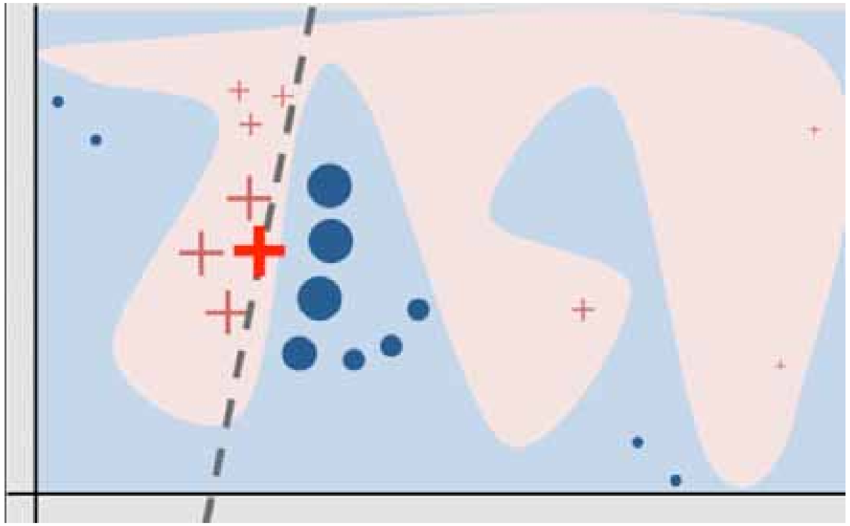
Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

- 00 Reflection
- 01 LIME – Local Interpretable Model Agnostic Explanations
- 02 BETA – Black Box Explanations through Transparent Approximation
- 03 LRP – Layer-wise Relevance Propagation
- 04 Deep Taylor Decomposition
- 05 Prediction Difference Analysis

01 LIME – Local Interpretable Model Agnostic Explanations

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

The central paper on 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 ...
☆ 99 Zitiert von: 2156 Ähnliche Artikel Alle 16 Versionen In EndNote importieren

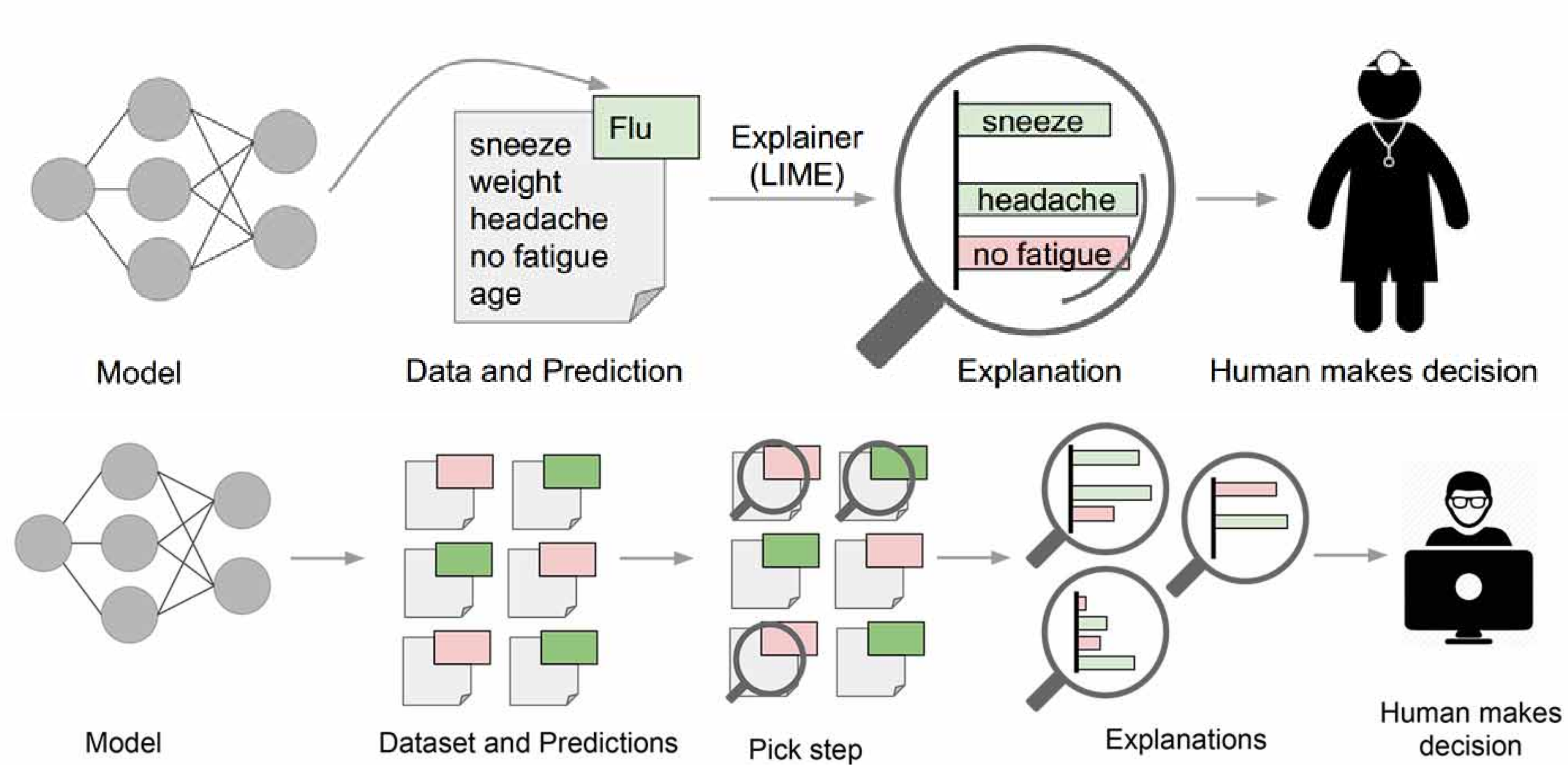
$$\xi(x) = \underset{g \in G}{\arg\min} \; L(f, g, \pi_x) + \Omega(g)$$

Fidelity score (for local fidelity)

Complexity score (for interpretability)

$\pi_x(z)$ Distance metric (in feature space!)

LIME Principle



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.

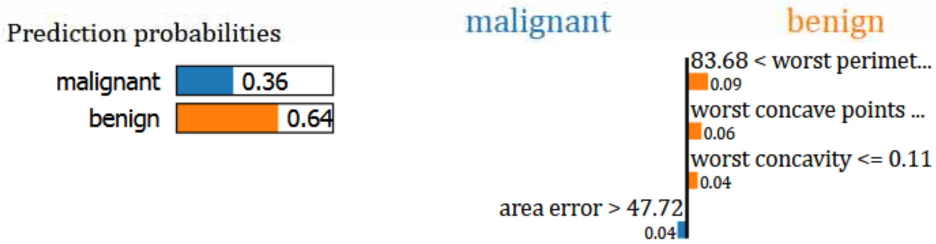
Example LIME – Model Agnostic Explanation

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)
```



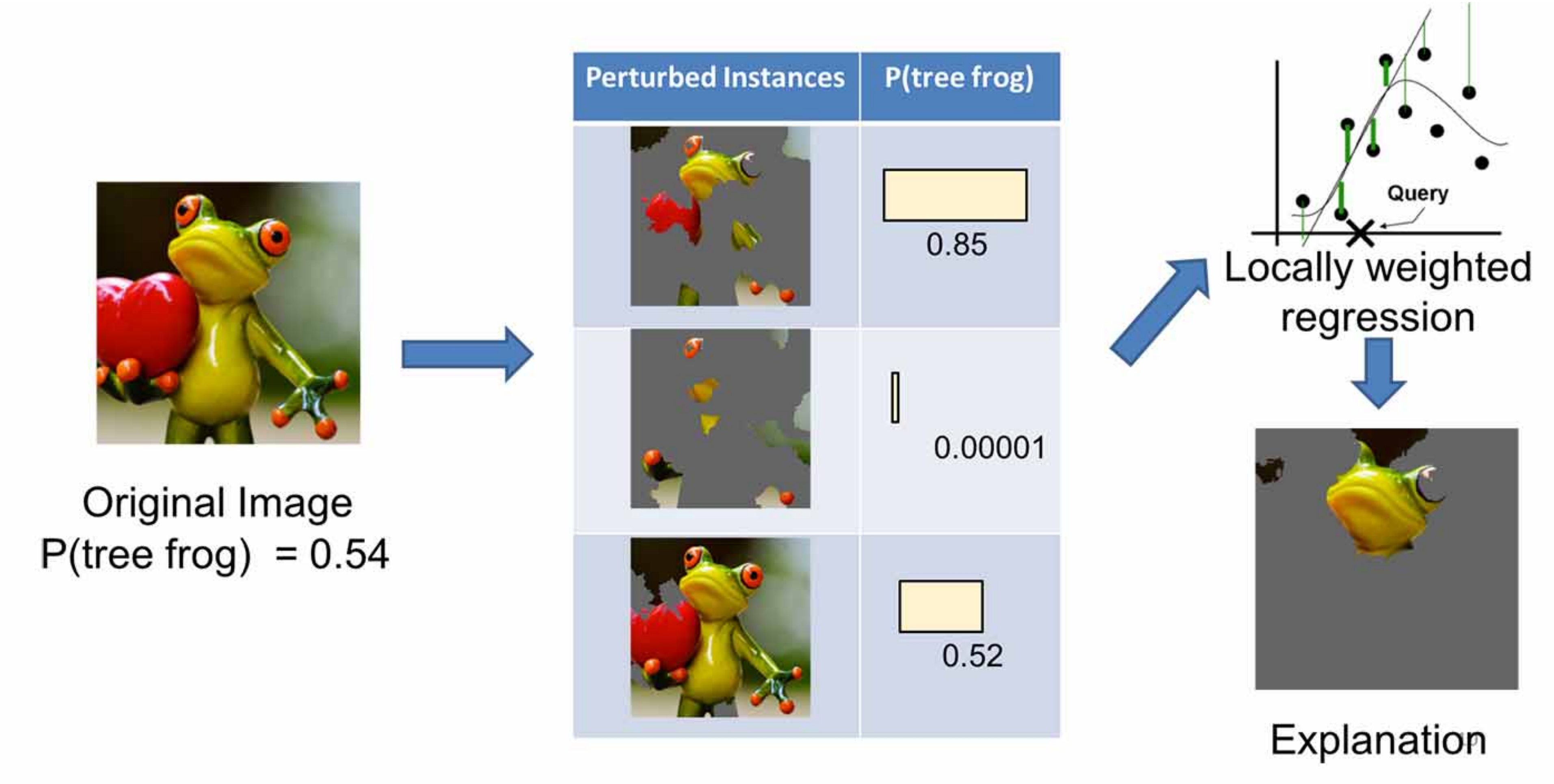
Feature	Value
worst perimeter	95.54
worst concave points	0.03
worst concavity	0.02
area error	106.00

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 malignant.

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.

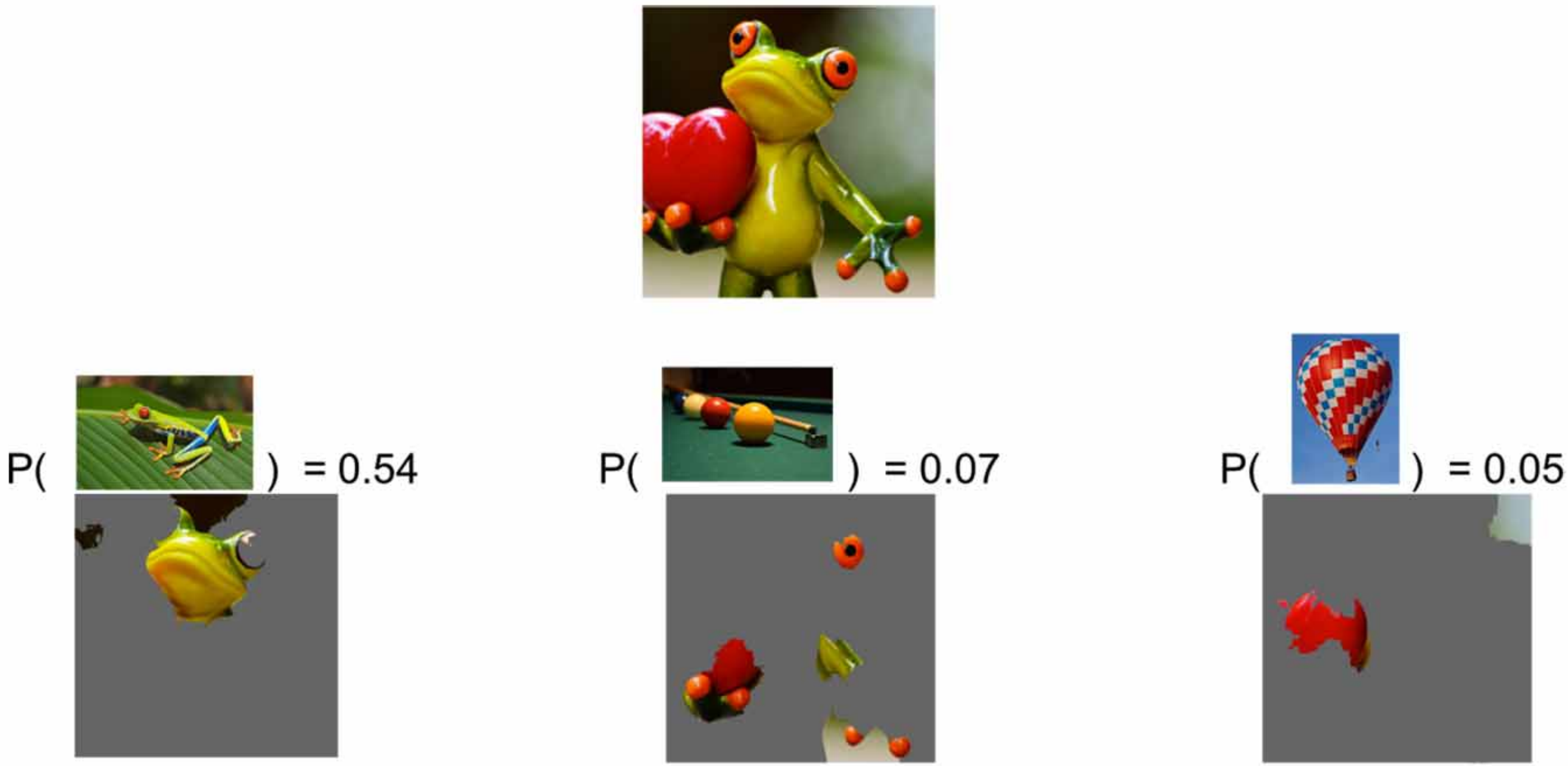
We chose 76 as it was a borderline decision. For example sample 86 is much more clear (this will we will set the `num_features` parameter to include all features so that we see each feature's contribution to the probability):

LIME Example



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

LIME Example



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

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

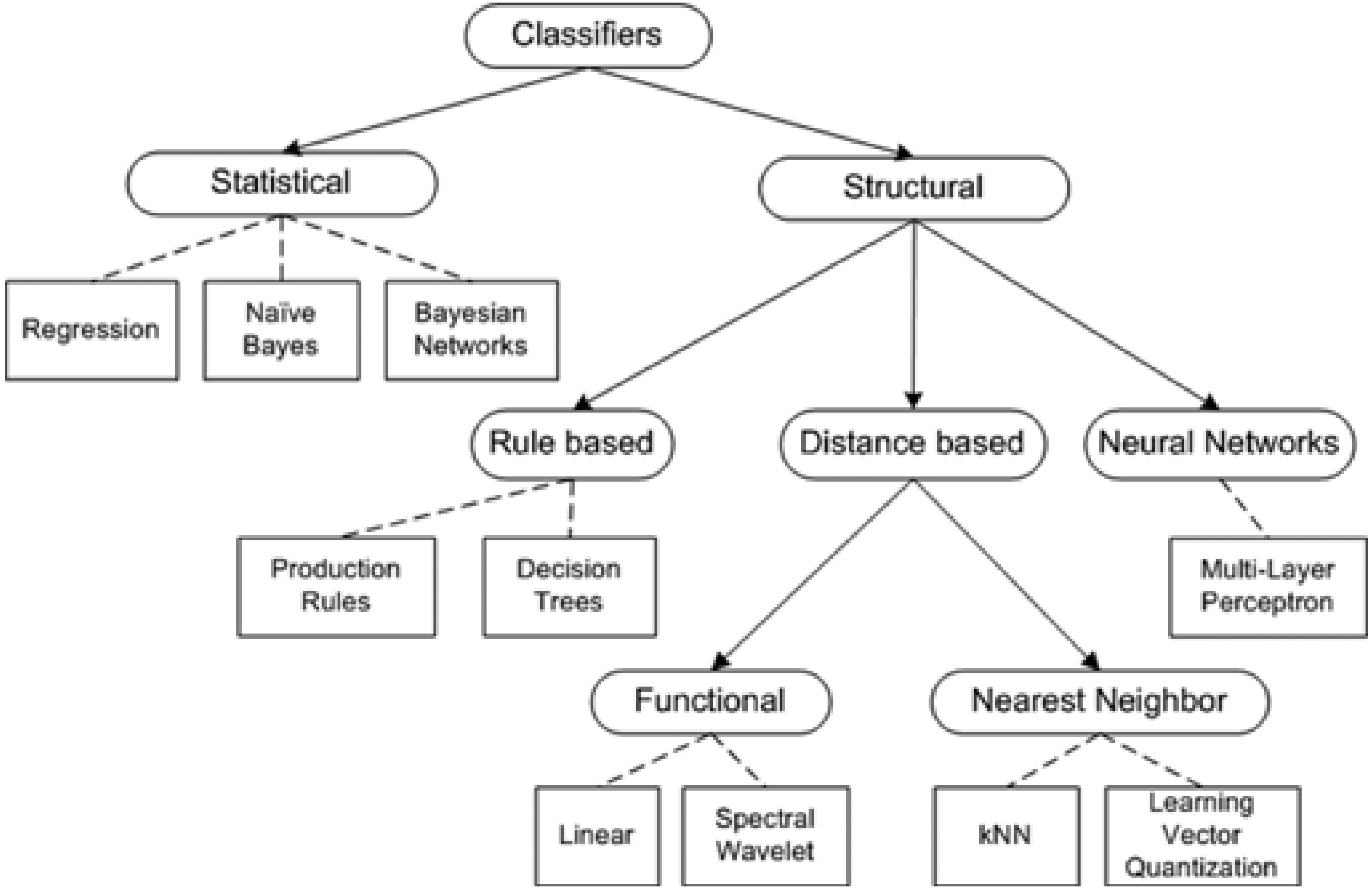
$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso($\mathcal{Z}, K$) $\triangleright$ with z'_i as features, $f(z)$ as target
return 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.

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

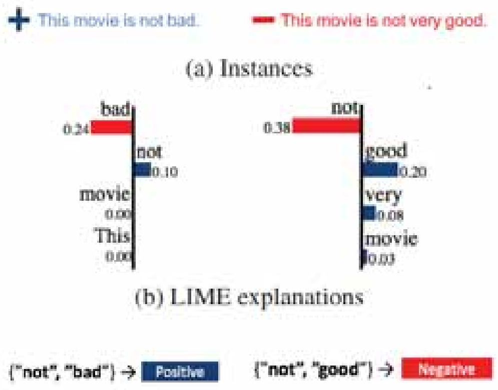
- - local model behaviour can be unrealistic
- - unclear coverage
- - ambiguity (how to select the kernel width)

Remember: there are myriads of classifiers …



https://stats.stackexchange.com/questions/271247/machine-learning-statistical-vs-structural-classifiers

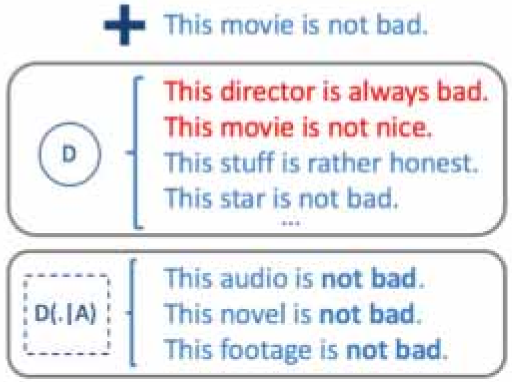
Follow-up: Anchor



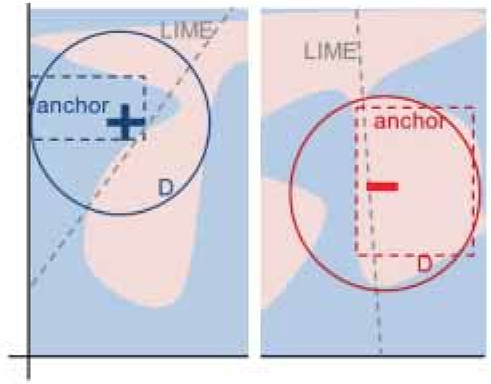
(a) Instances

(b) LIME explanations

(c) Anchor explanations



(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$



(b) Two toy visualizations

(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

What animal **is featured in this picture ?**	**dog**
What floor **is featured in this picture?**	dog
What toenail **is** paired **in this** flowchart **?**	dog
What animal **is** shown **on this** depiction **?**	dog

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the **dog**?	on the floor
What color is the **wall?**	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) **VQA:** More example anchors (in bold)

Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Anchors: High-precision model-agnostic explanations. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018 New Orleans. Association for the Advancement of Artificial Intelligence, 1527-1535.

More information

https://arteagac.github.io/blog.html

https://www.youtube.com/watch?v=vz_fkVkoGFM

https://www.youtube.com/watch?v=ENa-w65P1xM

https://www.youtube.com/watch?v=CY3t11vuuOM

02 BETA (Black Box Explanation through Transparent Approximation)

- 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

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

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 ...

☆ 99 Zitiert von: 214 Ähnliche Artikel Alle 10 Versionen In EndNote importieren

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 $\geq$ 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.

Optimization Process

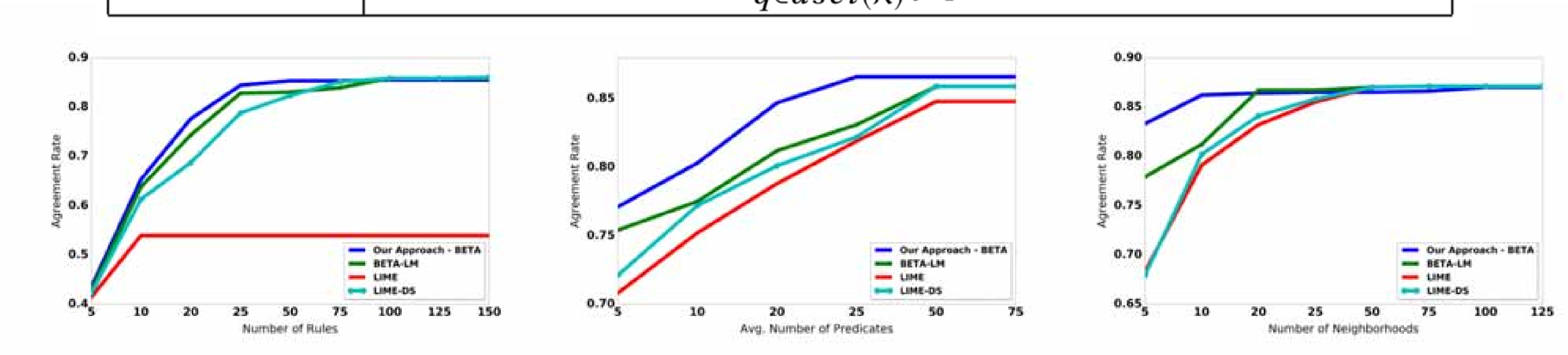
Algorithm 1 Optimization Procedure [5]

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 $f(S_i \setminus \{e\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \setminus e$
    Exchange Operation If $d \in X \setminus S_i$ and $e_j \in S_i$ (for $1 \leq j \leq k$) such that
    $(S_i \setminus e_j) \cup \{d\}$ (for $1 \leq j \leq k$) satisfies all the k constraints and
    $f(S_i \setminus \{e_1, e_2 \cdots e_k\} \cup \{d\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \setminus \{e_1, e_2, \cdots e_k\} \cup \{d\}$
  end while
  $V_{i+1} = V_i \setminus S_i$
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.

Measures of Fidelity, Interpretability, Unambiguity

Fidelity	$disagreement(\mathcal{R}) = \sum_{i=1}^{M} \lvert\{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}, \boldsymbol{x} \text{ satisfies } q_i \wedge s_i, \mathcal{B}(\boldsymbol{x}) \neq c_i\}\rvert$
Unambiguity	$ruleoverlap(\mathcal{R}) = \sum_{i=1}^{M} \sum_{j=1, i \neq j}^{M} overlap(q_i \wedge s_i, q_j \wedge s_j)$ $cover(\mathcal{R}) = \lvert\{x \mid x \in \mathcal{D}, x \text{ satisfies } q_i \wedge s_i \text{ where } i \in \{1 \cdots M\}\}\rvert$
Interpretability	$size(\mathcal{R})$: number of rules (triples of the form (q, s, c)) in $\mathcal{R}$ $maxwidth(\mathcal{R}) = \max_{e \in \bigcup_{i=1}^{M} (q_i \cup s_i)} width(e)$ $numpreds(\mathcal{R}) = \sum_{i=1}^{M} width(s_i) + width(q_i)$ $numdsets(\mathcal{R}) = \lvert dset(\mathcal{R})\rvert \text{ where } dset(\mathcal{R}) = \bigcup_{i=1}^{M} q_i$ $featureoverlap(\mathcal{R}) = \sum_{q \in dset(\mathcal{R})} \sum_{i=1}^{M} featureoverlap(q, s_i)$



BETA: Example of interpretable Decision set

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 ≥ 50	Predicate
s	Conjunction of one or more predicates, *e.g.*, Age ≥ 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

03 LRP (Layer-wise Relevance Propagation)

LRP at a glance

- 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.

LRP main papers

[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 …

[PDF] iNNvestigate neural networks!

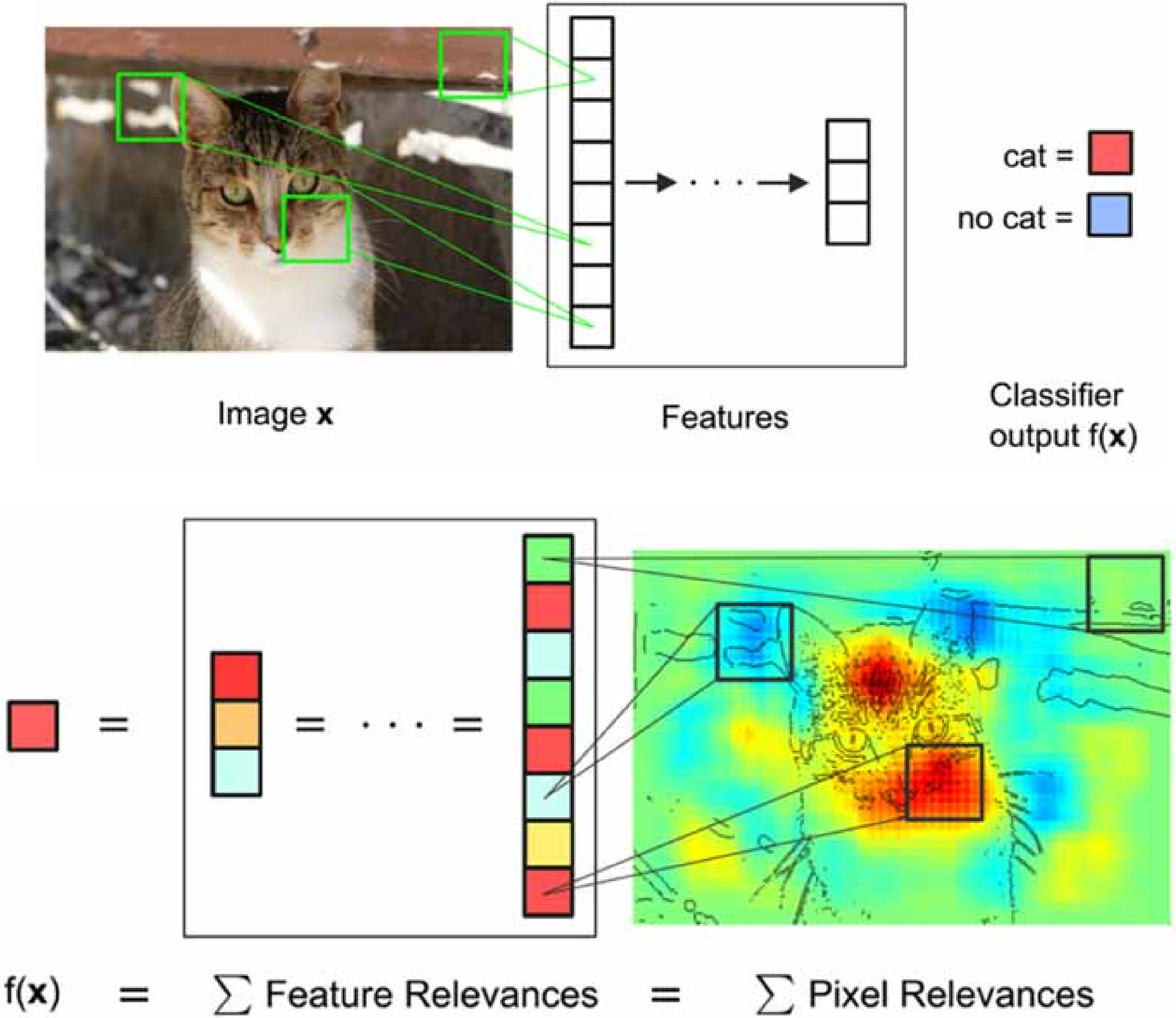
M Alber, S Lapuschkin, P Seegerer, M Hägele… - Journal of Machine …, 2019 - jmlr.org

… 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 …

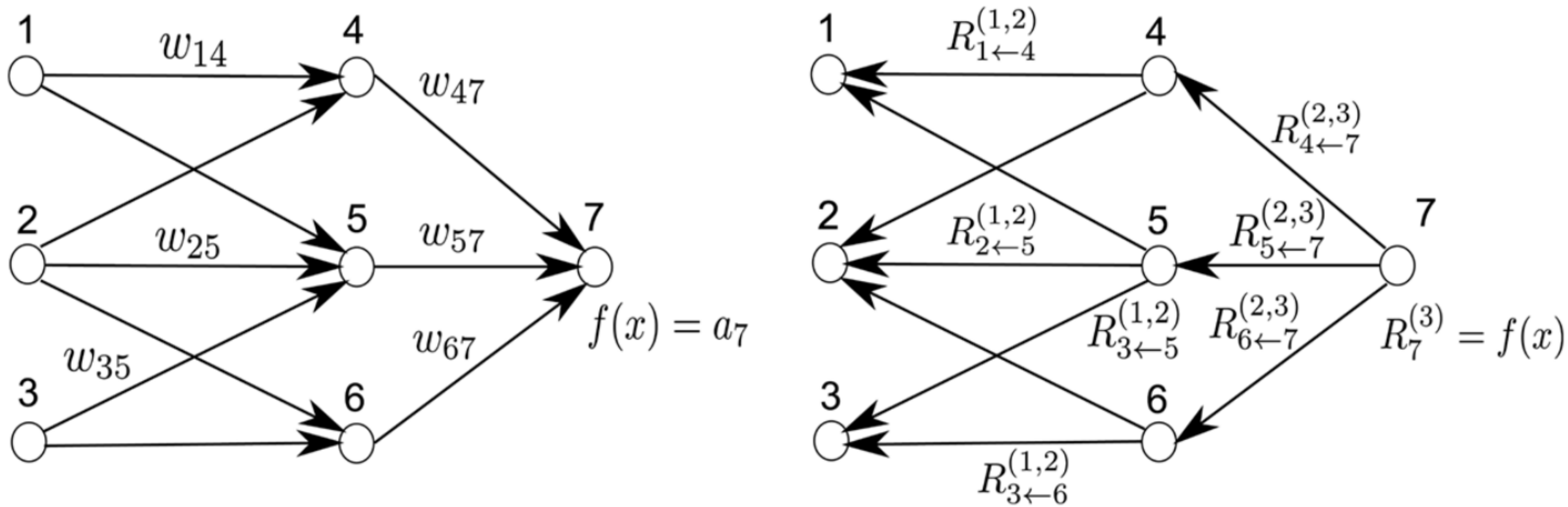
Grégoire Montavon 2019. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison. *In:* Samek, Wojciech, Montavon, Grégoire, Vedaldi, Andrea, Hansen, Lars Kai & Müller, Klaus-Robert (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Cham: Springer International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

LRP Layer-Wise Relevance Propagation

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.



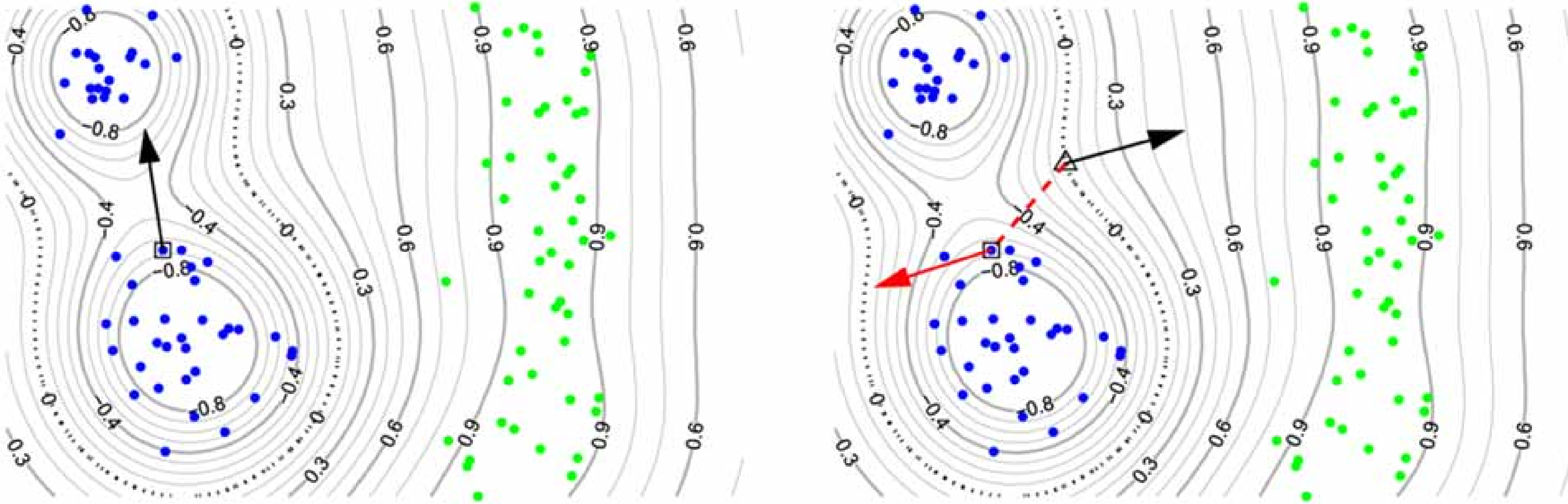
A NN-classifier during prediction time



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

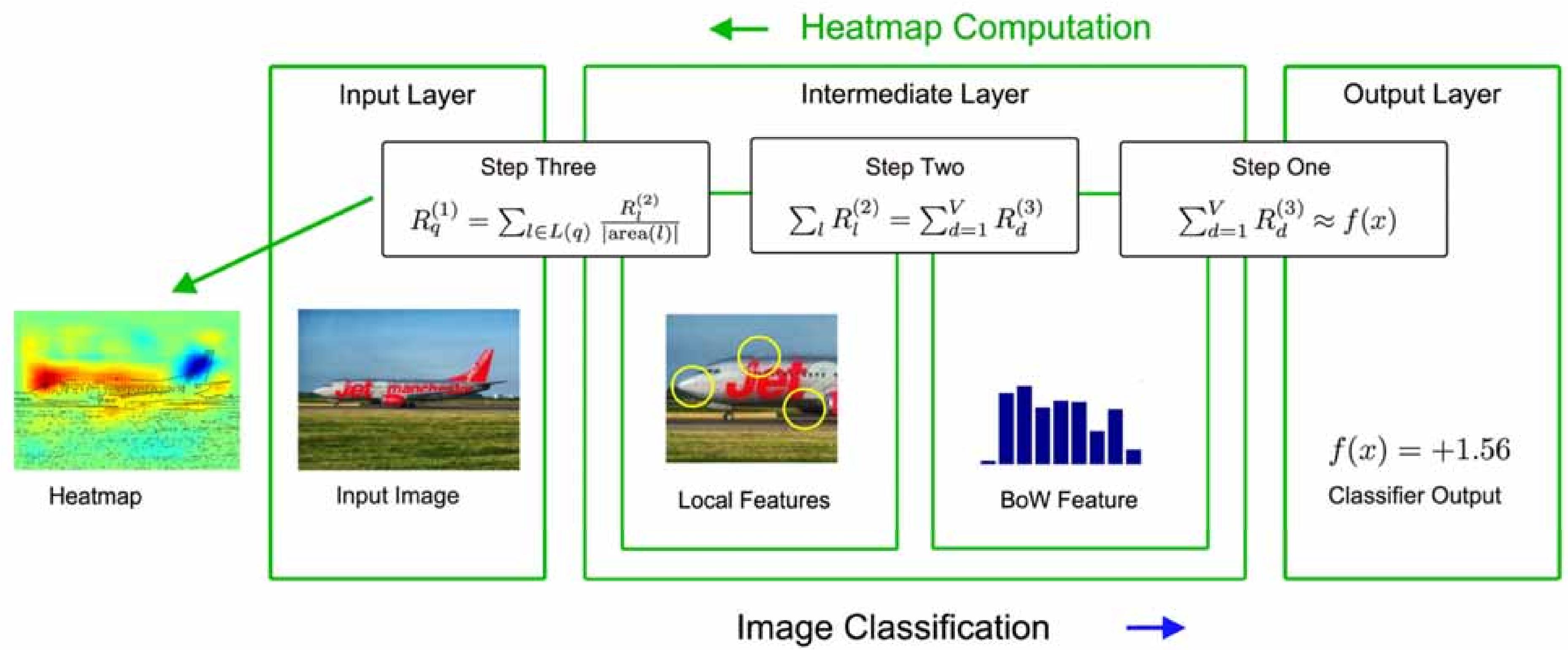
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

Example Taylor Decomposition



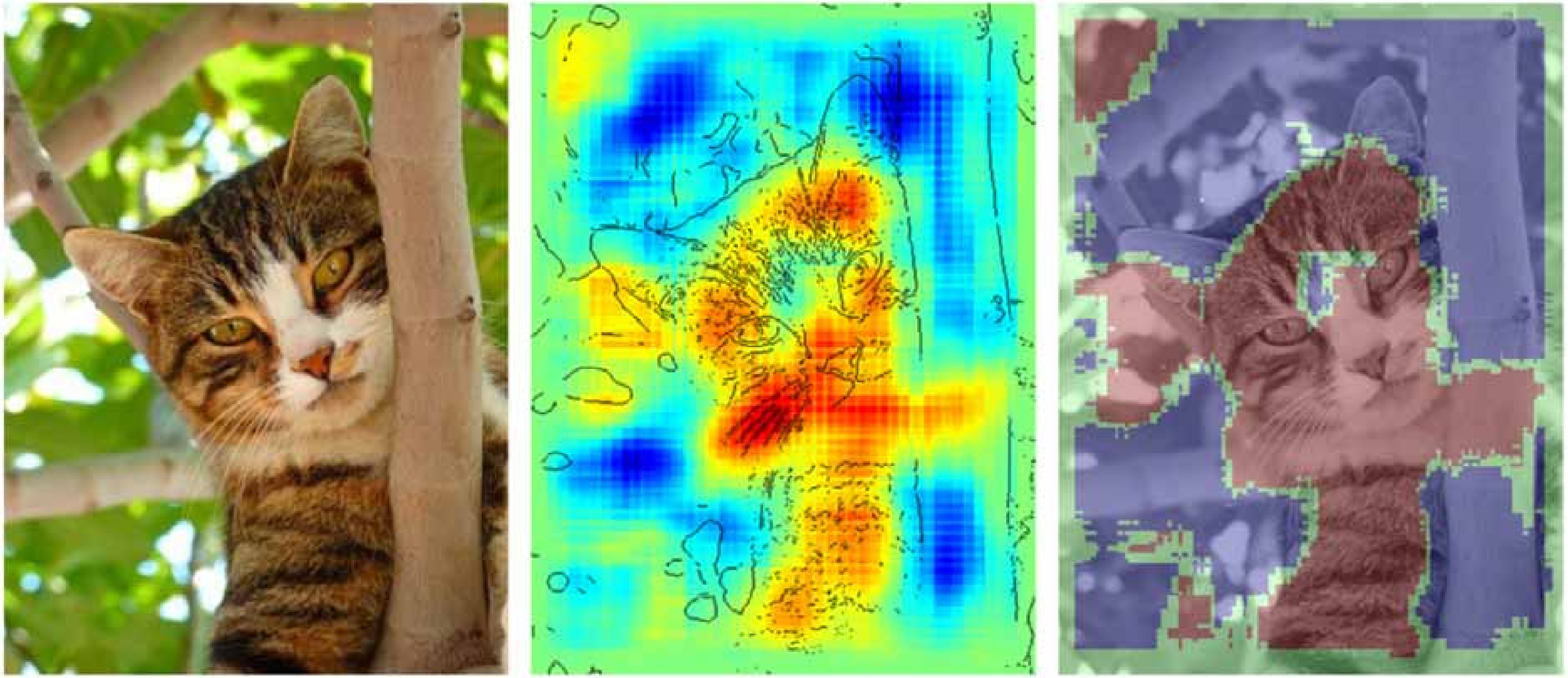
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

Heatmap Computation



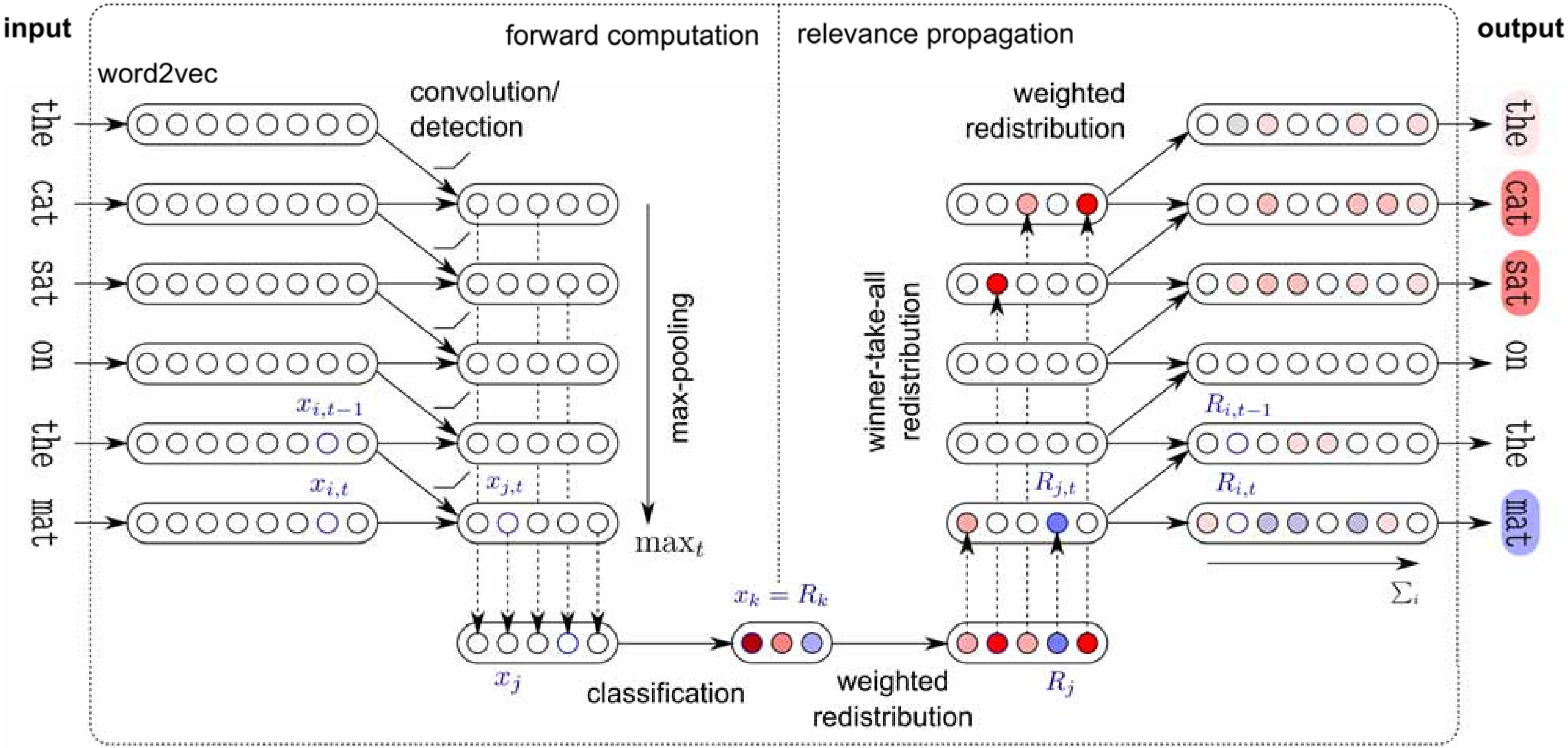
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

Pixel-wise decomposition for bag-of-words features



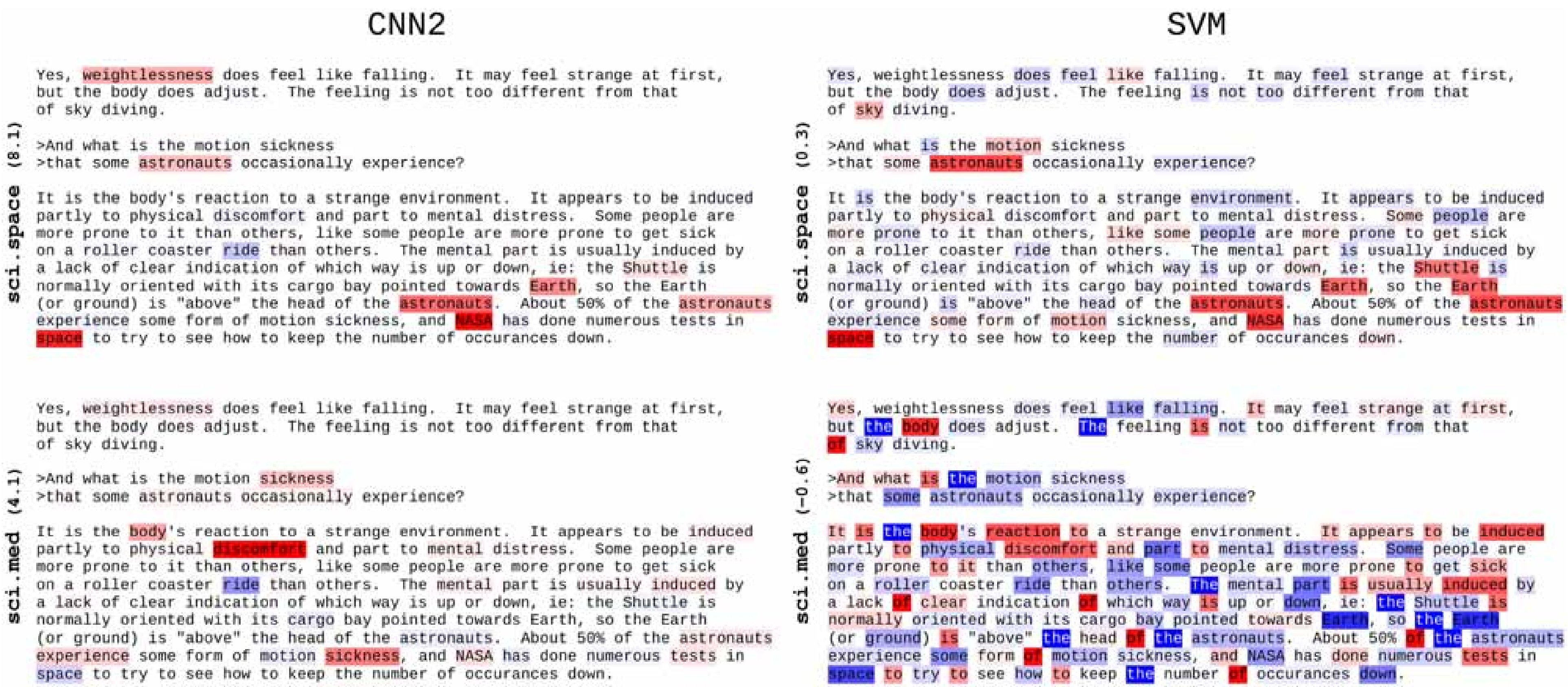
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

What is relevant in a text document?



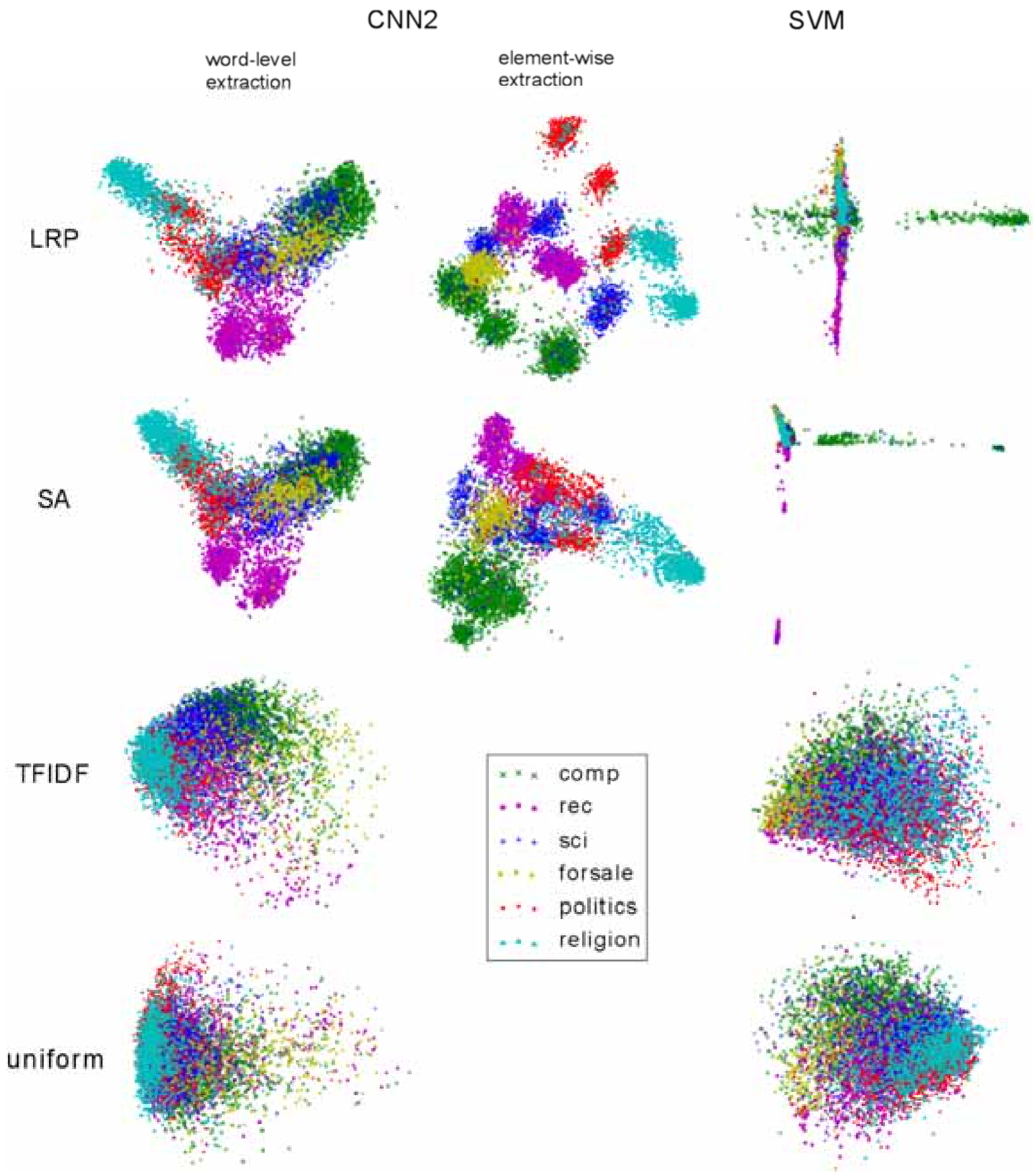
Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller & Wojciech Samek 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12, (8), e0181142, doi:10.1371/journal.pone.0181142.

Example: What is relevant in a text document?



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.

PCA-Projection of the summary vectors



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.

LRP on GitHub

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)

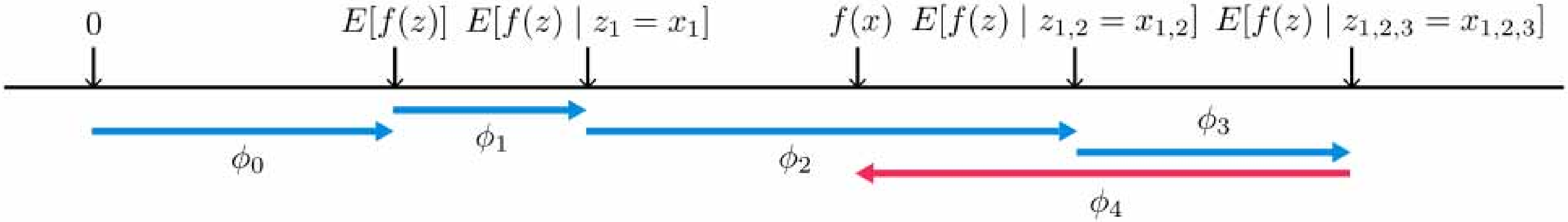
https://github.com/albermax/innvestigate

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

Alernatively: (SHapley Additive exPlanations)



Theorem 2 (Shapley kernel) *Under Definition 1, the specific forms of $\pi_{x'}$, L, and Ω that make solutions of Equation 2 consistent with Properties 1 through 3 are:*

$$\Omega(g) = 0,$$
$$\pi_{x'}(z') = \frac{(M-1)}{(M\ choose\ |z'|)|z'|(M-|z'|)},$$
$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x(z')) - g(z')\right]^2 \pi_{x'}(z'),$$

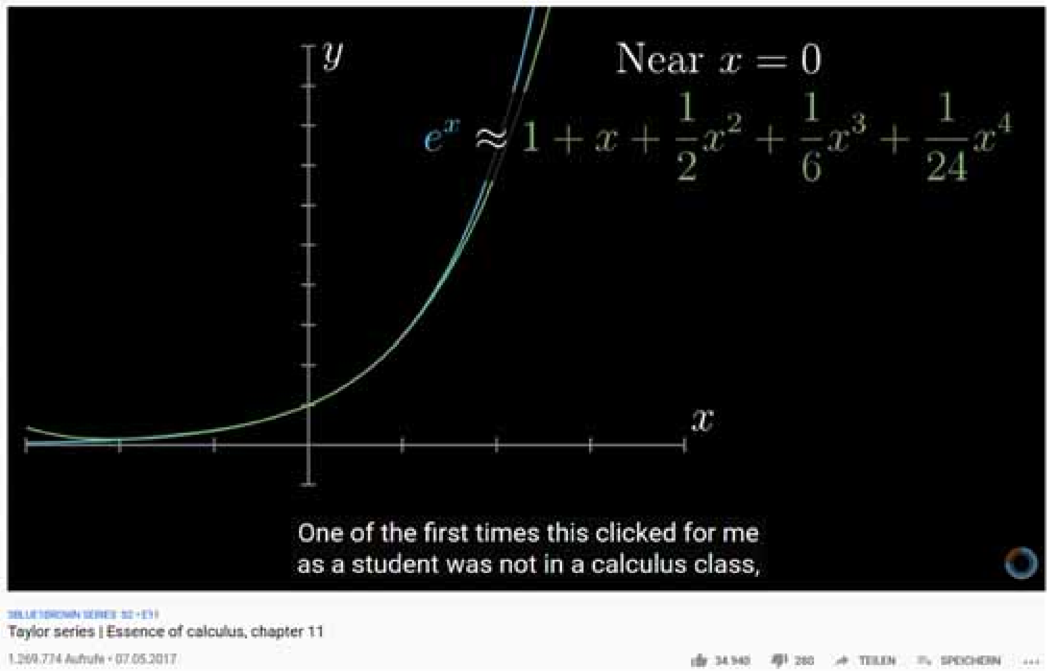
where $|z'|$ is the number of non-zero elements in z'.

Scott M. Lundberg & Su-In Lee. A unified approach to interpreting model predictions. *In:* Guyon, Isabelle, Luxburg, Ulrike Von, Bengio, Samy, Wallach, Hanna, Fergus, Rob, Viswanathan, Svn & Garnett, Roman, eds. Advances in Neural Information Processing Systems, 2017 Montreal. NIPS, 4765-4774.

https://github.com/OpenXAIProject/PyConKorea2019-Tutorials

04 Deep Taylor Decomposition

Remember: Taylor Series



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

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

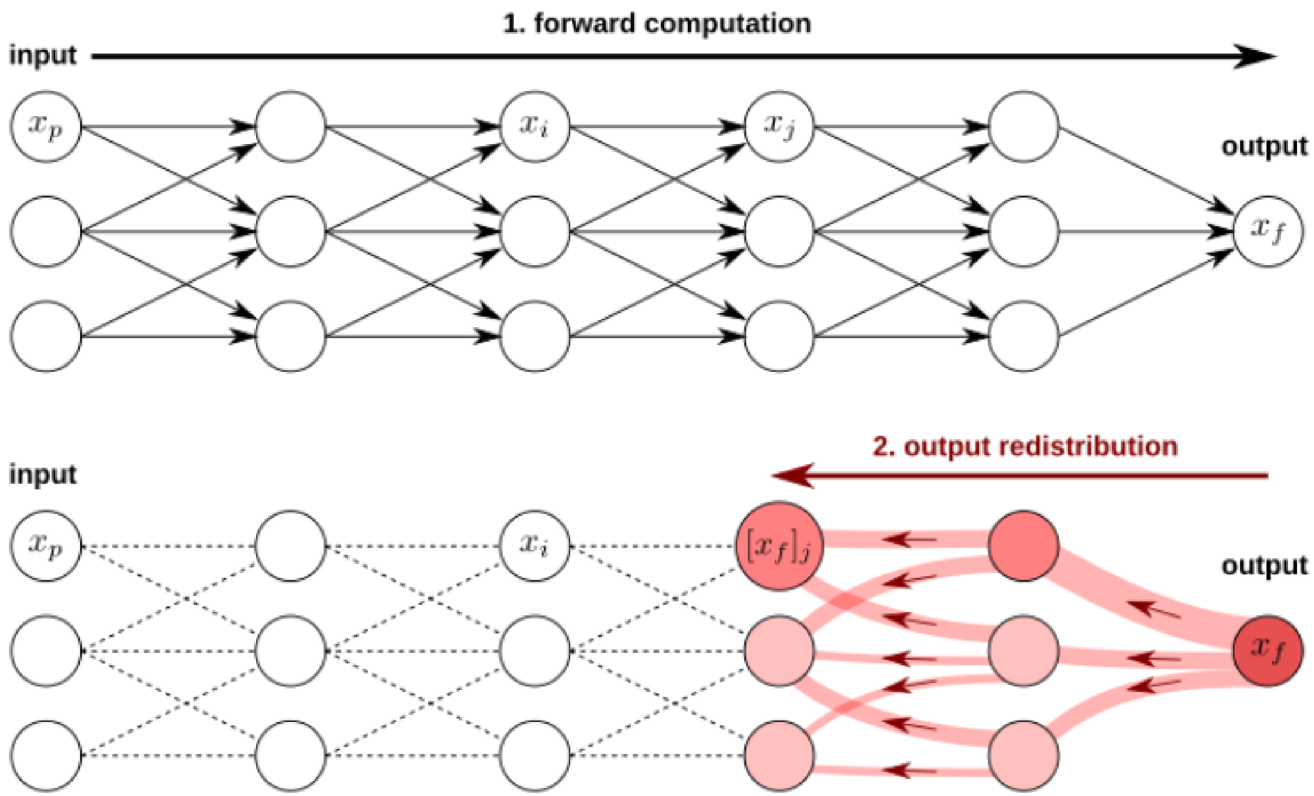
Brook Taylor

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

- 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

Definitions

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:

$$\forall\, \boldsymbol{x}: f(\boldsymbol{x}) = \sum_p R_p(\boldsymbol{x}).$$

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

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

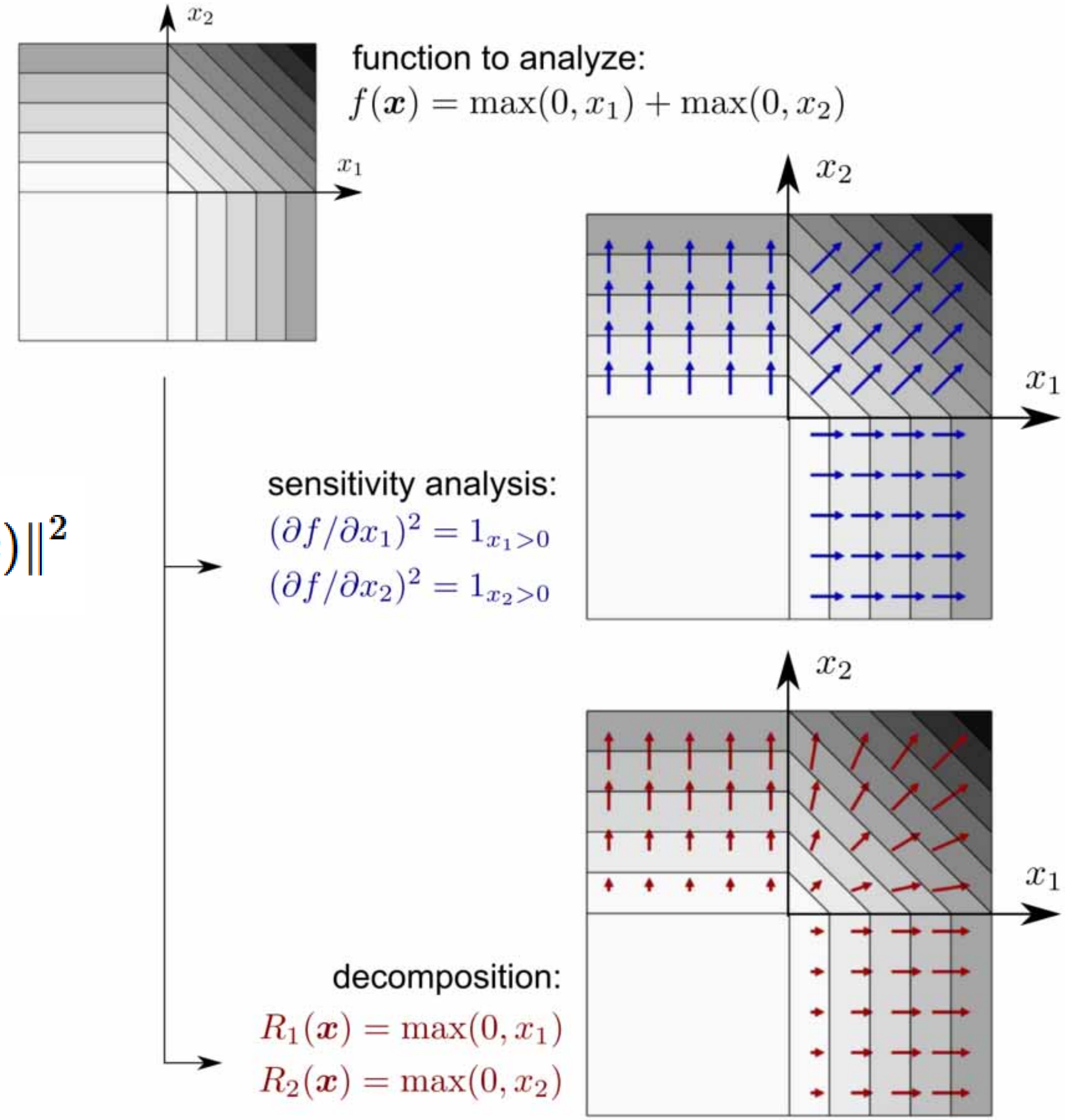
Definition 3. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *consistent* if it is conservative *and* positive. That is, it is consistent if it complies with Definitions 1 and 2.

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.

Sensitivity Analysis vs. Decomposition

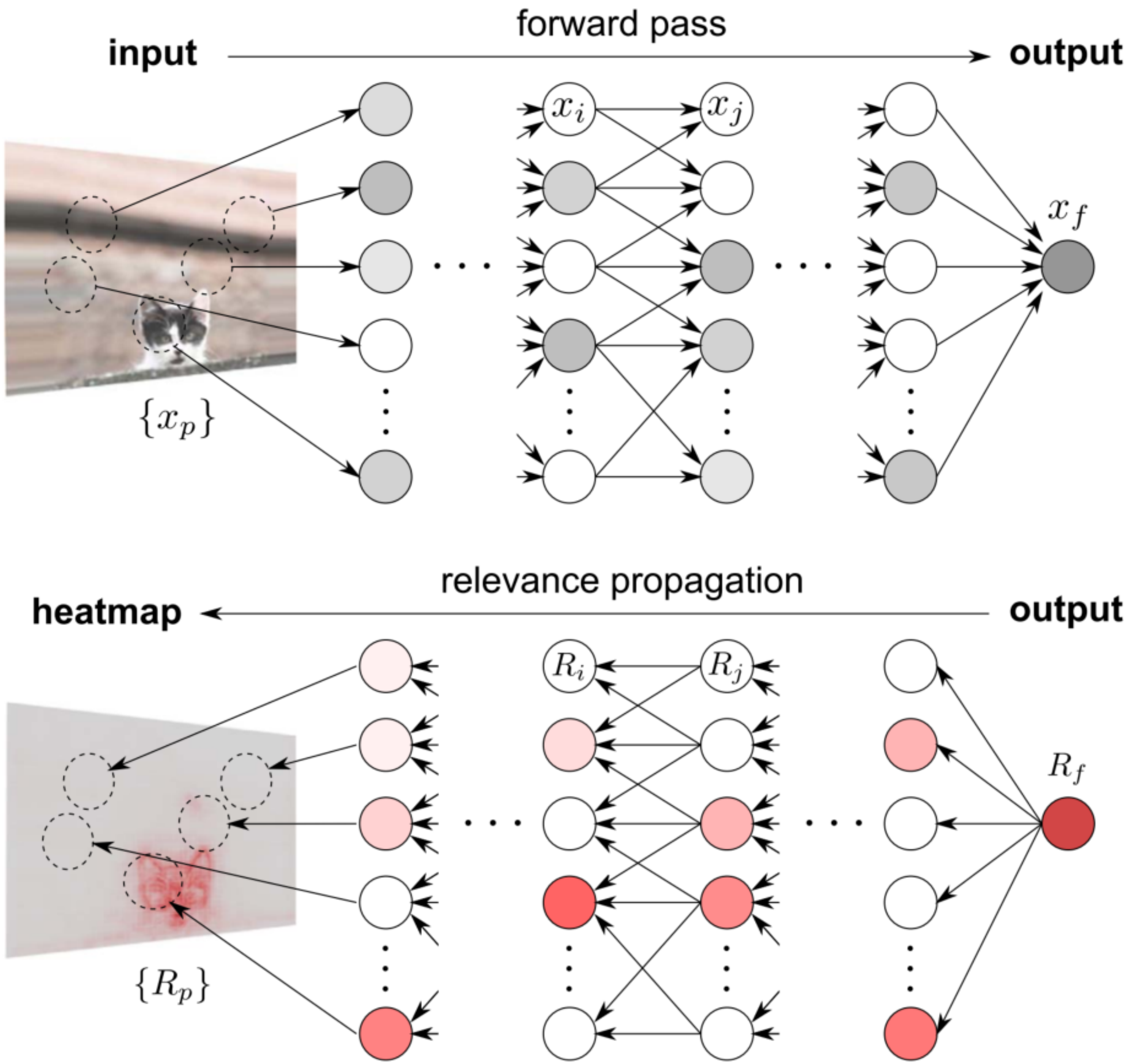
$$\sum_p (\partial f / \partial x_p)^2 = \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|^2$$

$$\sum_p [f(\boldsymbol{x})]_p = f(\boldsymbol{x})$$

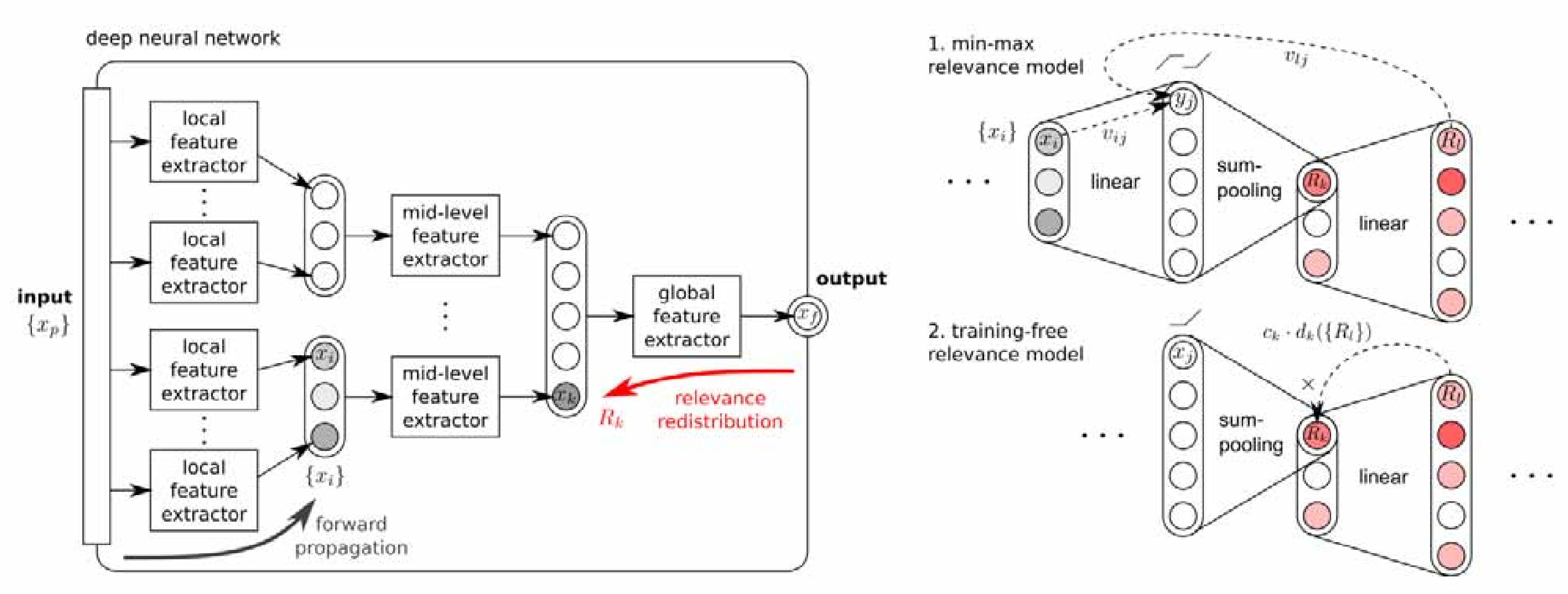


Relevance propagation

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.

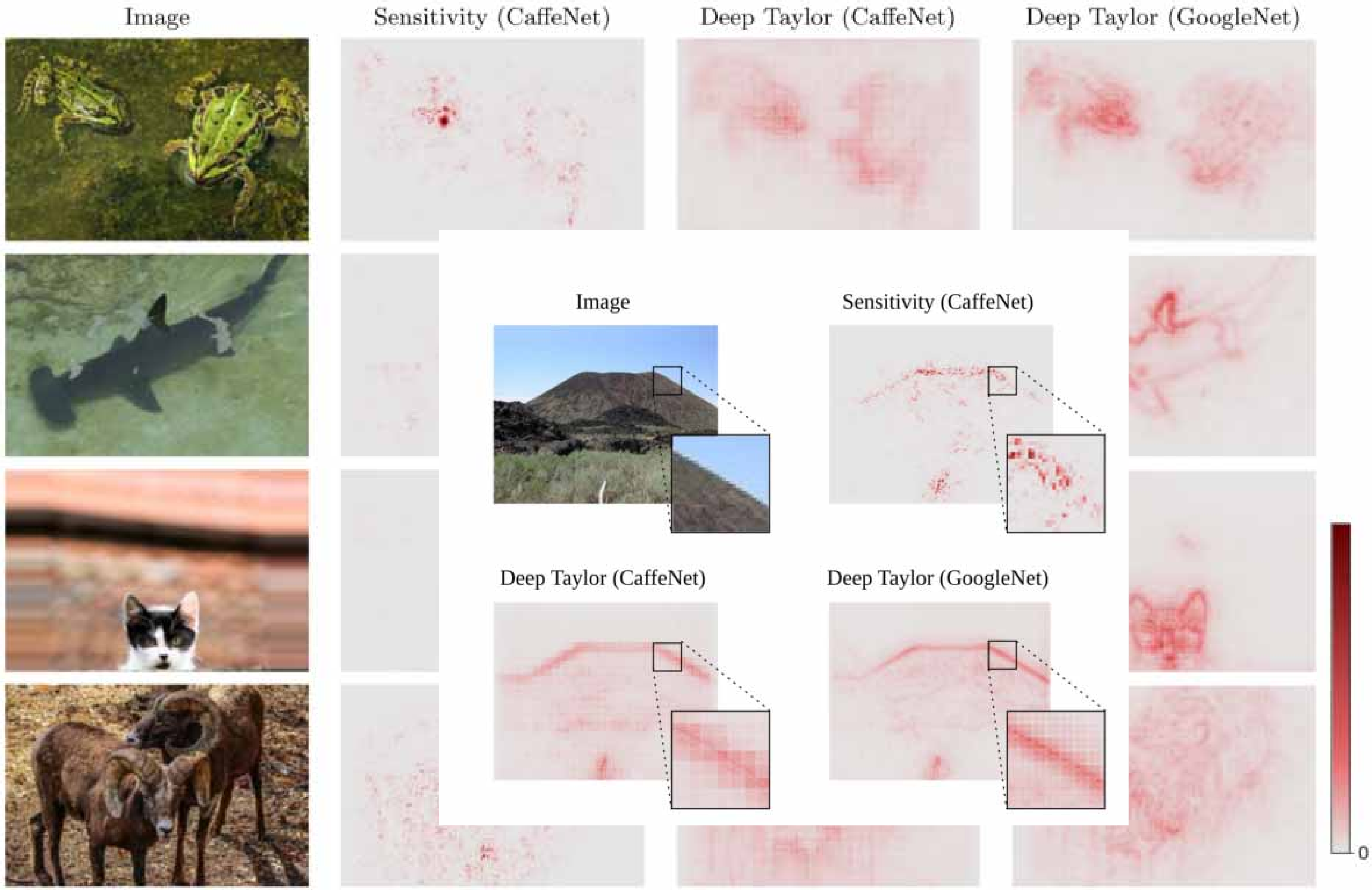


Relevance Redistribution

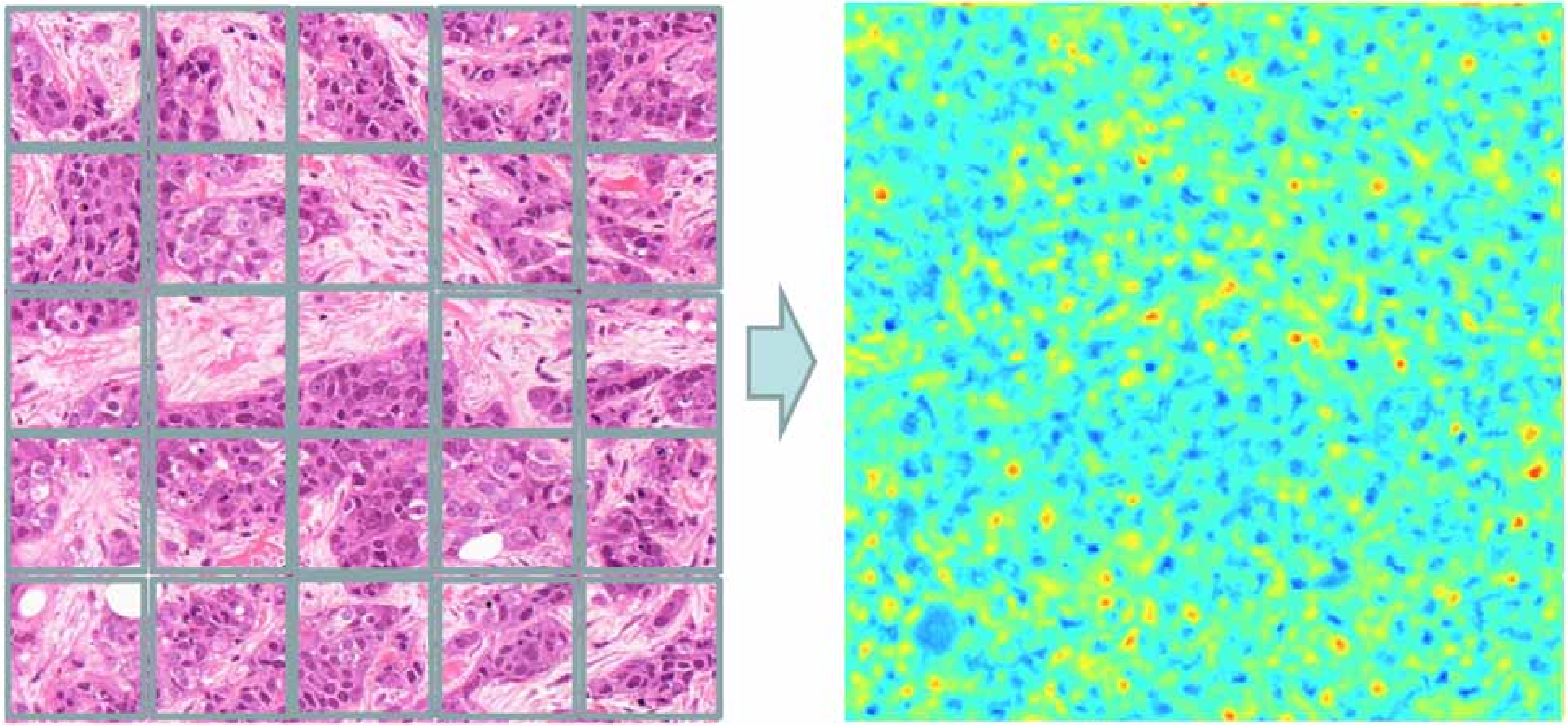


Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.

Example 1: Comparison



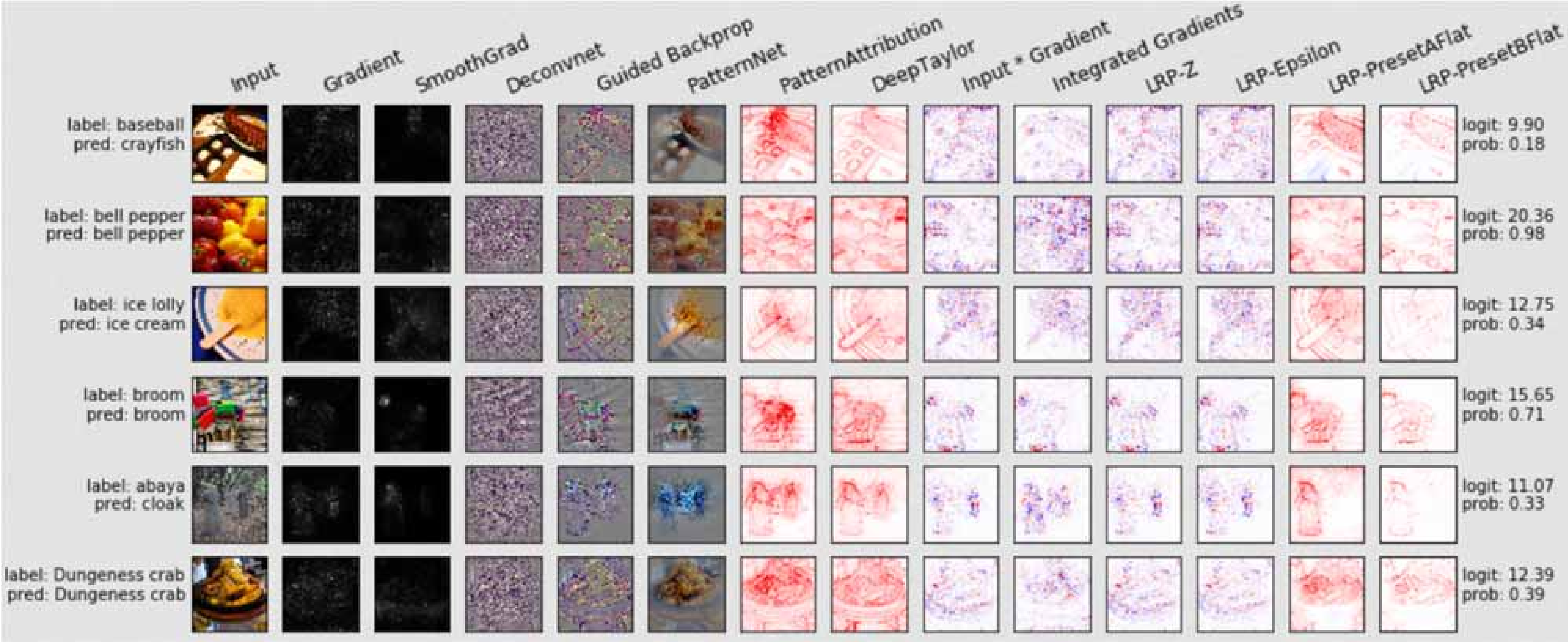
Example 2 Histopathology



Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller & Alexander Binder 2019. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *arXiv:1908.06943*.

Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Albrecht Stenzinger, Laura Parlow, Jan Budczies & Benjamin Goeppert 2018. Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. *arXiv:1805.11178vl*.

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek & Sebastian Lapuschkin 2019. Towards best practice in explaining neural network decisions with LRP. *arXiv:1910.09840*.



https://github.com/albermax/innvestigate

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

05 Prediction Difference Analysis

Core Paper

Visualizing deep neural network decisions: Prediction difference analysis

LM Zintgraf, TS Cohen, T Adel, M Welling - arXiv preprint arXiv …, 2017 - arxiv.org

This article presents the prediction difference analysis method for visualizing the response of a deep neural network to a specific input. When classifying images, the method highlights areas in a given input image that provide evidence for or against a certain class. It …

Relevance Value

$$p(c|\mathbf{x}_{\setminus i}) = \sum_{x_i} p(x_i|\mathbf{x}_{\setminus i}) p(c|\mathbf{x}_{\setminus i}, x_i)$$

$$p(c|\mathbf{x}_{\setminus i}) \approx \sum_{x_i} p(x_i) p(c|\mathbf{x}_{\setminus i}, x_i)$$

$$\mathrm{WE}_i(c|\mathbf{x}) = \log_2\left(\mathrm{odds}(c|\mathbf{x})\right) - \log_2\left(\mathrm{odds}(c|\mathbf{x}_{\setminus i})\right)$$

Marko Robnik-Šikonja & Igor Kononenko 2008. Explaining Classifications For Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, 20, (5), 589-600, doi:10.1109/TKDE.2007.190734.

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel & Max Welling 2017. Visualizing deep neural network decisions: Prediction difference analysis. arXiv:1702.04595.

https://github.com/lmzintgraf/DeepVis-PredDiff/blob/master/README.md

https://openreview.net/forum?id=BJ5UeU9xx

Evaluating the prediction difference

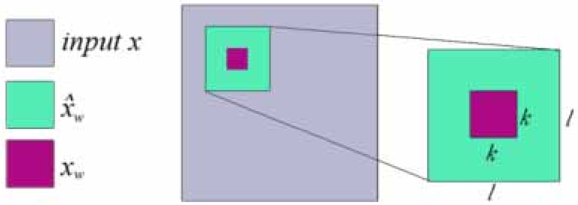


Figure 2: Simple illustration of the sampling procedure in algorithm 1. Given the input image x, we select every possible patch x_w (in a sliding window fashion) of size $k \times k$ and place a larger patch $\hat{x}_w$ of size $l \times l$ around it. We can then conditionally sample x_w by conditioning on the surrounding patch $\hat{x}_w$.

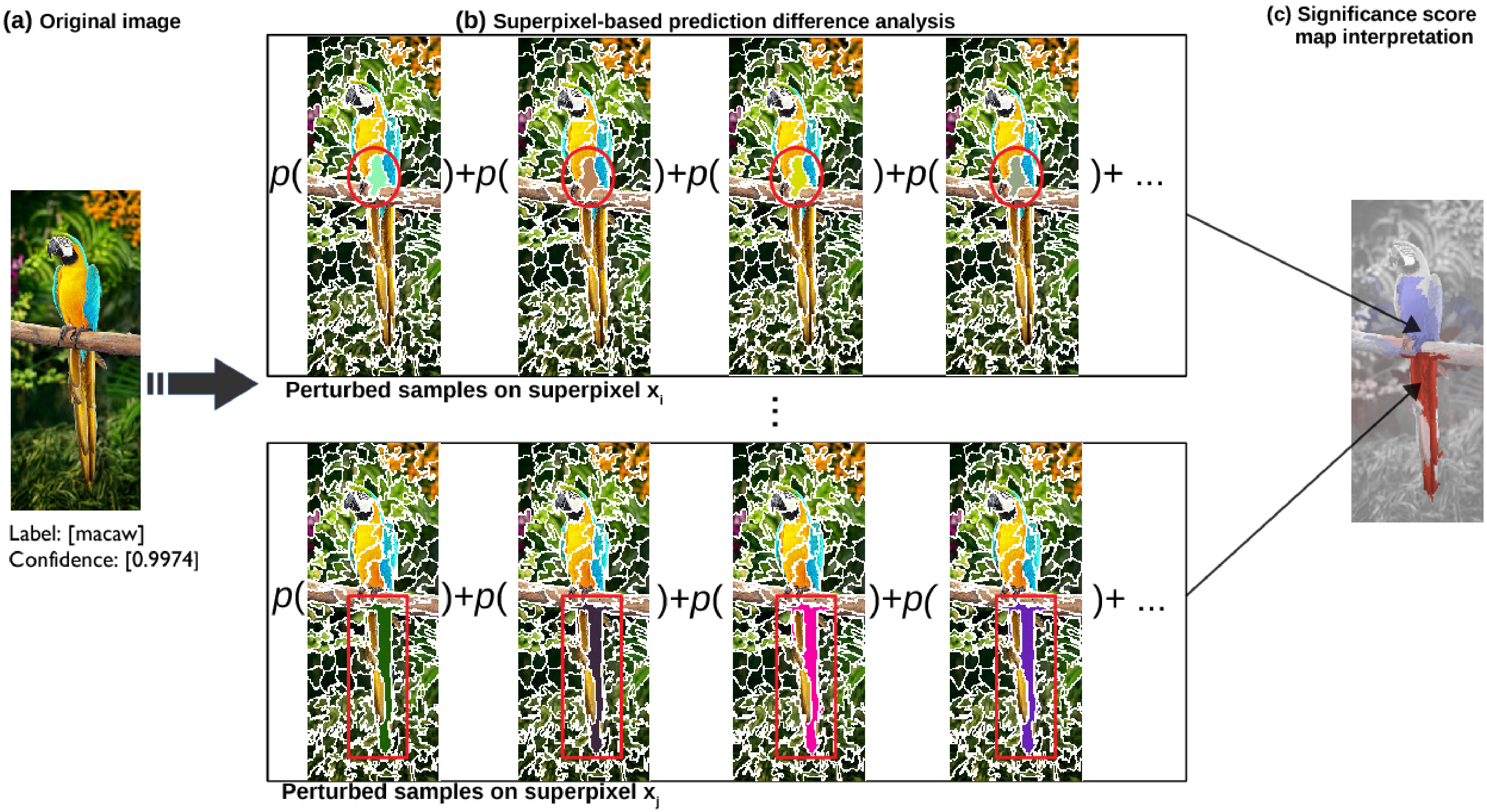
Algorithm 1 Evaluating the prediction difference using conditional and multivariate sampling

---

Input: classifier with outputs p(c|x), input image $\mathbf{x}$ of size $n \times n$, inner patch size k, outer patch size $l > k$, class of interest c, probabilistic model over patches of size $l \times l$, number of samples S
Initialization: WE = zeros(n*n), counts = zeros(n*n)
for every patch $\mathbf{x}_w$ of size $k \times k$ **in x do**
  $\mathbf{x}' = \text{copy}(\mathbf{x})$
  $\text{sum}_w = 0$
  define patch $\hat{\mathbf{x}}_w$ of size $l \times l$ that contains $\mathbf{x}_w$
  for $s = 1$ **to** S **do**
    $\mathbf{x}'_w \leftarrow \mathbf{x}_w$ sampled from $p(\mathbf{x}_w | \hat{\mathbf{x}}_w \backslash \mathbf{x}_w)$
    $\text{sum}_w \mathrel{+}= p(c|\mathbf{x}')$ ▷ evaluate classifier
  end for
  $p(c|\mathbf{x}\backslash\mathbf{x}_w) := \text{sum}_w / S$
  WE[coordinates of $\mathbf{x}_w$] $\mathrel{+}= \log_2(\text{odds}(c|\mathbf{x})) - \log_2(\text{odds}(c|\mathbf{x}\backslash\mathbf{x}_w))$
  counts[coordinates of $\mathbf{x}_w$] $\mathrel{+}= 1$
end for
Output: WE / counts ▷ point-wise division

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel & Max Welling 2017. Visualizing deep neural network decisions: Prediction difference analysis. arXiv:1702.04595.

Superpixel-based prediction difference analysis



Yi Wei, Ming-Ching Chang, Yiming Ying, Ser Nam Lim & Siwei Lyu. Explain Black-box Image Classifications Using Superpixel-based Interpretation. 2018 24th International Conference on Pattern Recognition (ICPR), 2018. IEEE, 1640-1645.

Contextual Prediction Difference Analysis

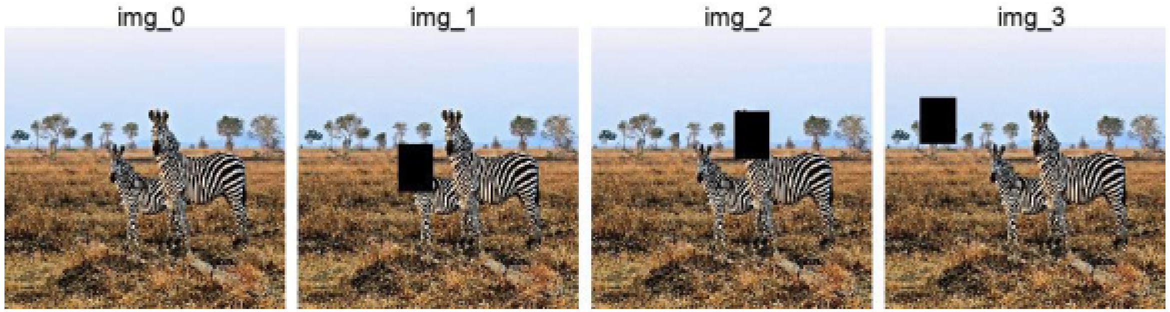


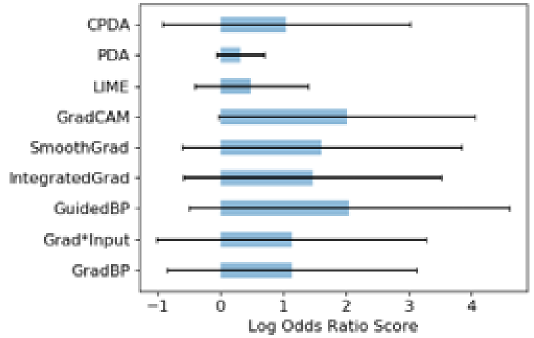
Image	img_0	img_1	img_2	img_3
Alexnet	.9999	.9999(+.0000)	.9999(+.0000)	.9999(+.0000)
VGG16	.9998	.9995(-.0003)	.9997(-.0001)	.9997(-.0001)
Inception_V3	.9415	.9365(-0.050)	.9587(+.0172)	.9336(-.0079)
ResNet	.9945	.9983(+.0038)	.9986(+.0041)	.9964(+.0019)
DenseNet	.9817	.9920(+.0103)	.9712(-.0105)	.9803(-.0012)

$$r_i = \sum_{j \neq i} \frac{R_{\backslash j}}{|R_{\backslash j}|}$$

$$r_1 = \frac{R_{\backslash 2}}{|R_{\backslash 2}|} + \frac{R_{\backslash 3}}{|R_{\backslash 3}|} = 0/2 + 1/2 = 0.5$$

$$r_2 = \frac{R_{\backslash 1}}{|R_{\backslash 1}|} + \frac{R_{\backslash 3}}{|R_{\backslash 3}|} = 0/2 + 1/2 = 0.5$$

$$r_3 = \frac{R_{\backslash 1}}{|R_{\backslash 1}|} + \frac{R_{\backslash 2}}{|R_{\backslash 2}|} = 0/2 + 0/2 = 0$$



$$R_{\backslash i} = f(\boldsymbol{x}) - \sum_{k=1}^{M} p(\boldsymbol{x}_{\backslash i} = \boldsymbol{v}_k | x_i) p(y | x_i, \boldsymbol{x}_{\backslash i} = \boldsymbol{v}_k) = f(\boldsymbol{x}) - p(y|x_i) = f(\boldsymbol{x}) - f(x_i)$$

Jindong Gu & Volker Tresp 2019. Contextual Prediction Difference Analysis. *arXiv:1910.09086*.

Thank you!