Slides and
Resources

Post Hoc, Local and Model-Agnostic Explanations

AAAI 2023 Tutorial on Trustworthy and Responsible AI

Yilun Zhou

MIT CSAIL & Amazon

Outline

Slides and Resources

- Why model explanations?
- How to compute model explanations? -- Definitions
- How to evaluate model explanations? -- Evaluations
- A definition-evaluation duality

Why Model Explanations?

Slides and Resources

Why Model Explanations?

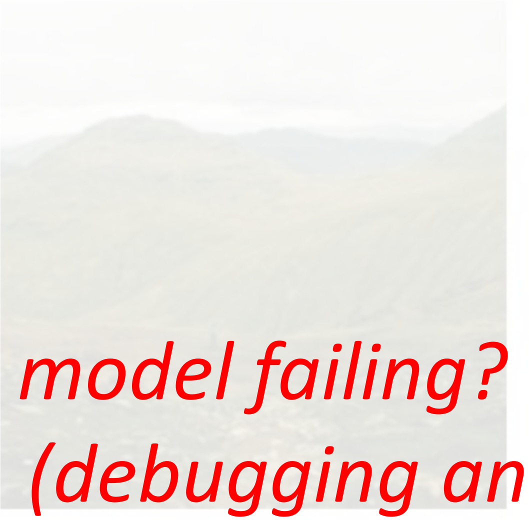
Slides and Resources

Task for DNN	Caption image	Recognise pneumonia
Problem	Describes green hillside as grazing sheep	Fails on scans from new hospitals
Shortcut	Uses background to recognise primary object	Looks at hospital token, not lung

(Geirhos et al. 2020)

Why Model Explanations?

Slides and Resources



Why is my model failing?

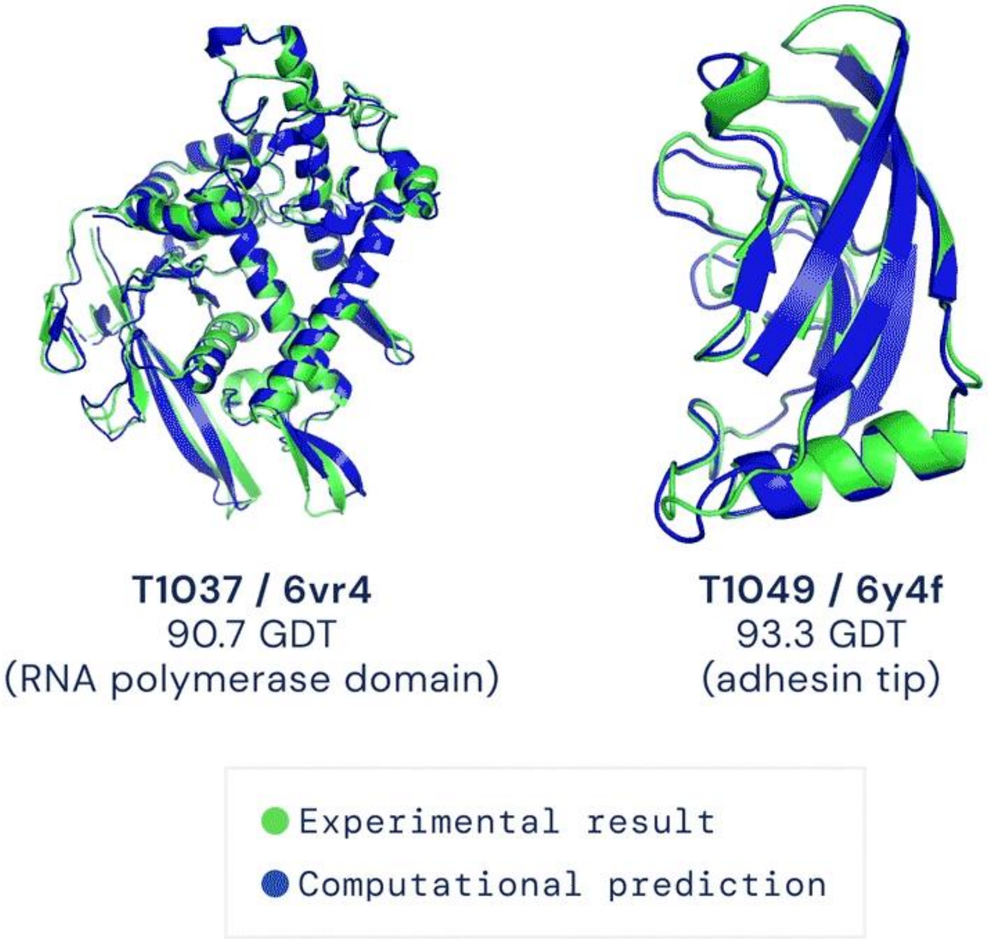
Because … (debugging and diagnosis)

Task for DNN	Caption image	Recognise pneumonia
Problem	Describes green hillside as grazing sheep	Fails on scans from new hospitals
Shortcut	Uses background to recognise primary object	Looks at hospital token, not lung

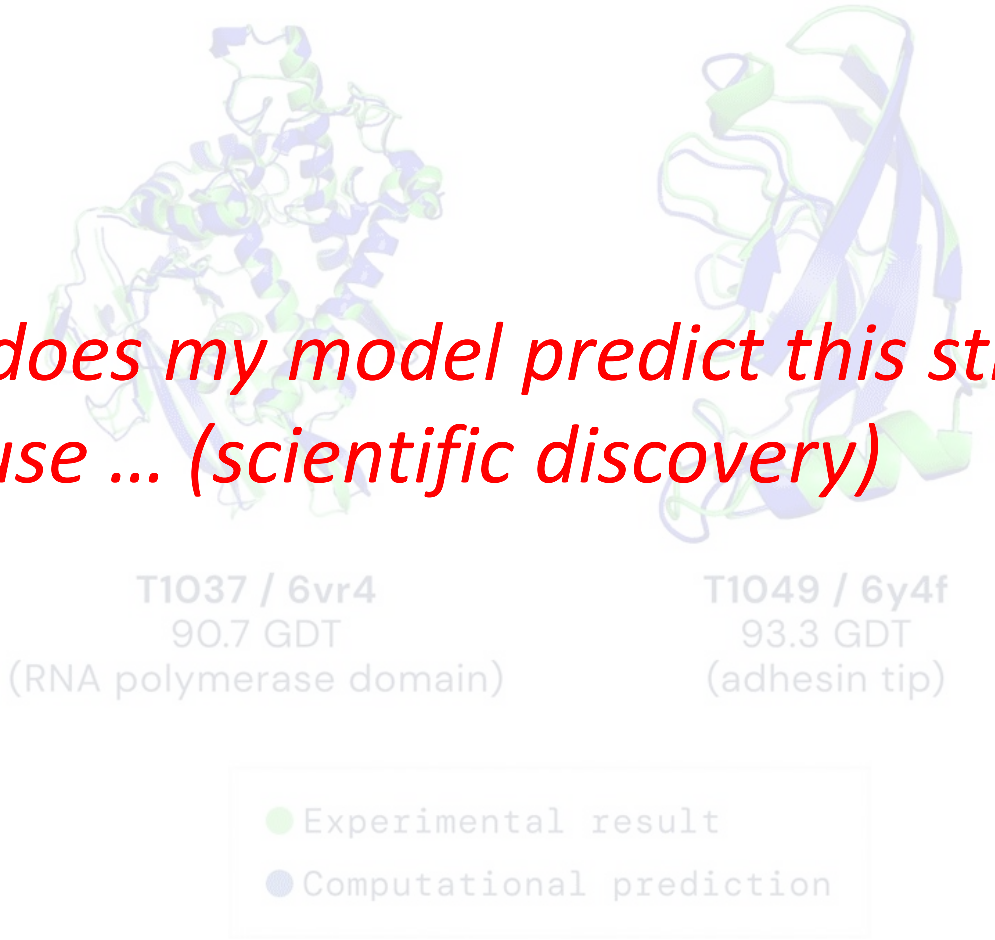
(Geirhos et al. 2020)

Why Model Explanations?

Slides and Resources



Why Model Explanations?



How does my model predict this structure?

Because … (scientific discovery)

The Two Major Axes of Interpretability

Slides and Resources

Global
Explains the model "more generally"

Local
Requires an input instance

The Two Major Axes of Interpretability

Slides and Resources

Post Hoc

Generated by an external explainer after model prediction

Global

Explains the model “more generally”

Local

Requires an input instance

Intrinsic

Generated during model prediction

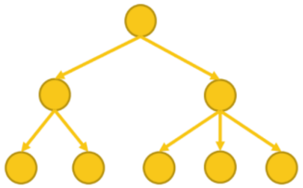
Slides and Resources

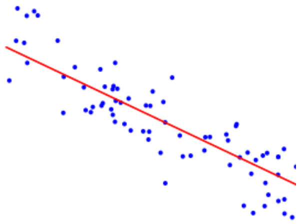
The Two Major Axes of Interpretability

Post Hoc
Generated by an external explainer after model prediction

Global
Explains the model “more generally”

Local
Requires an input instance





Intrinsic
Generated during model prediction

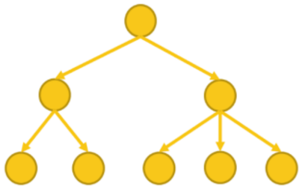
The Two Major Axes of Interpretability

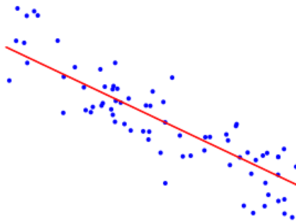
Slides and Resources

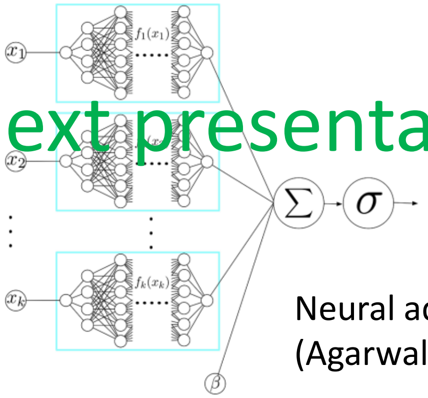
Post Hoc
Generated by an external explainer after model prediction

Global
Explains the model "more generally"

Local
Requires an input instance







Next presentation

Neural additive model
(Agarwal et al., 2021)

Intrinsic
Generated during model prediction

The Two Major Axes of Interpretability

Slides and Resources

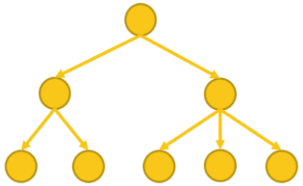
CNN filter visualization
(Olah et al., 2017)

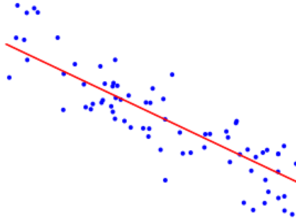
Post Hoc
Generated by an external explainer after model prediction

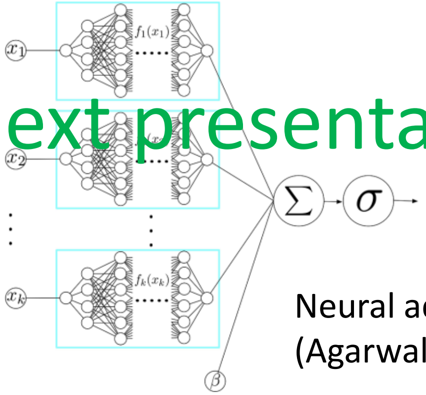
Global
Explains the model “more generally”

Local
Requires an input instance

Next presentation







Neural additive model
(Agarwal et al., 2021)

Intrinsic
Generated during model prediction

The Two Major Axes of Interpretability

Slides and Resources

Post Hoc

Generated by an external explainer after model prediction

CNN filter visualization
(Olah et al., 2017)

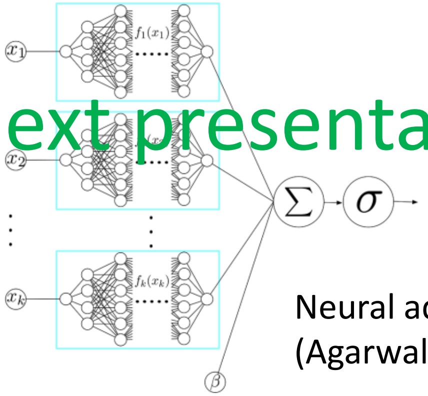
This presentation

Global

Explains the model “more generally”

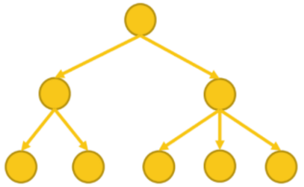
Local

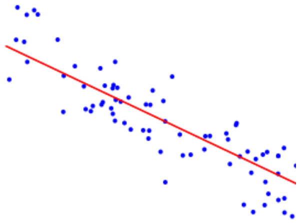
Requires an input instance



Next presentation

Neural additive model
(Agarwal et al., 2021)





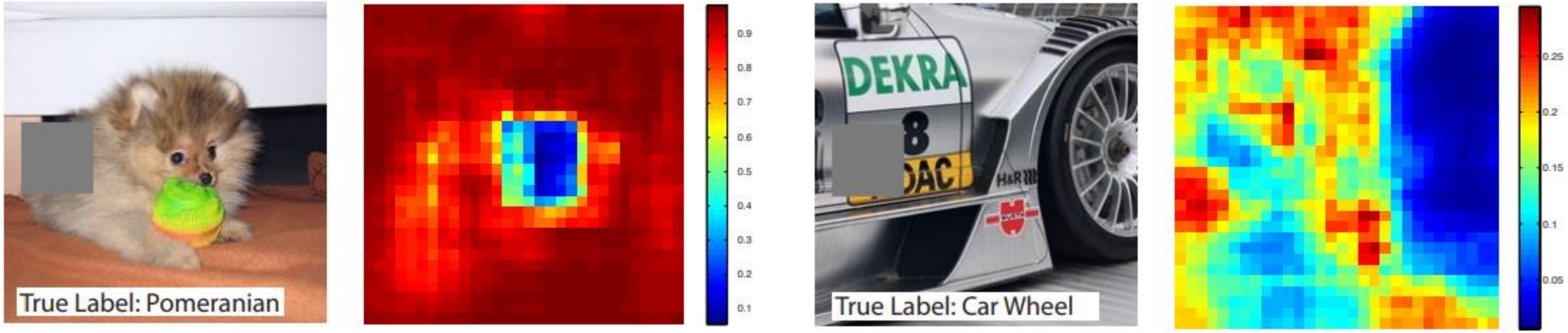
Intrinsic

Generated during model prediction

Slides and Resources

Local *Post Hoc* Explanations

- What is a local *post hoc* explanation?
  - Some description of the model's local decision making logic
- If this image patch is grayed out, the model prediction will change to this.



Occlusion Saliency (Zeiler and Fergus, 2014)

Local *Post Hoc* Explanations

Slides and Resources

- What is a local *post hoc* explanation?
  - Some description of the model's local decision making logic
- If this input feature is changed to this value, the model prediction will be different.

Mortgage Application
- Age: 30
- Job: Manager
- Salary: $50,000
- Debt: $5,000
- Marital Status: Single
- House Price: $500K
- Decision: Denied

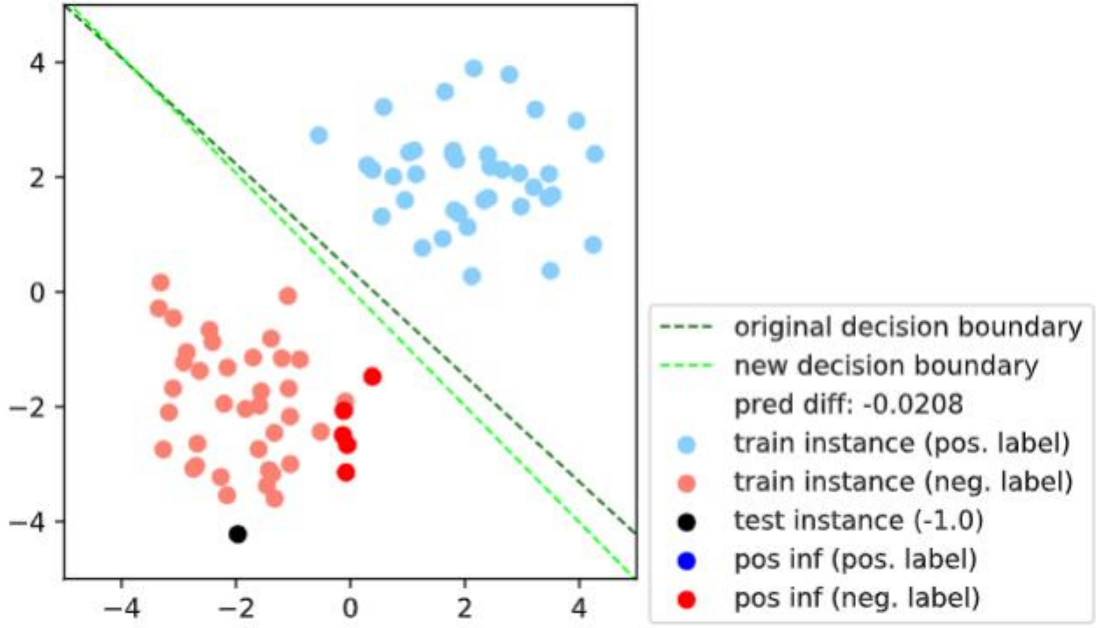
Counterfactual Explanation →

Mortgage Application
- Age: 30
- Job: Manager
- Salary: ~~$50,000~~ $60,000
- Debt: $5,000
- Marital Status: Single
- House Price: $500K
- Decision: Approved

Local *Post Hoc* Explanations

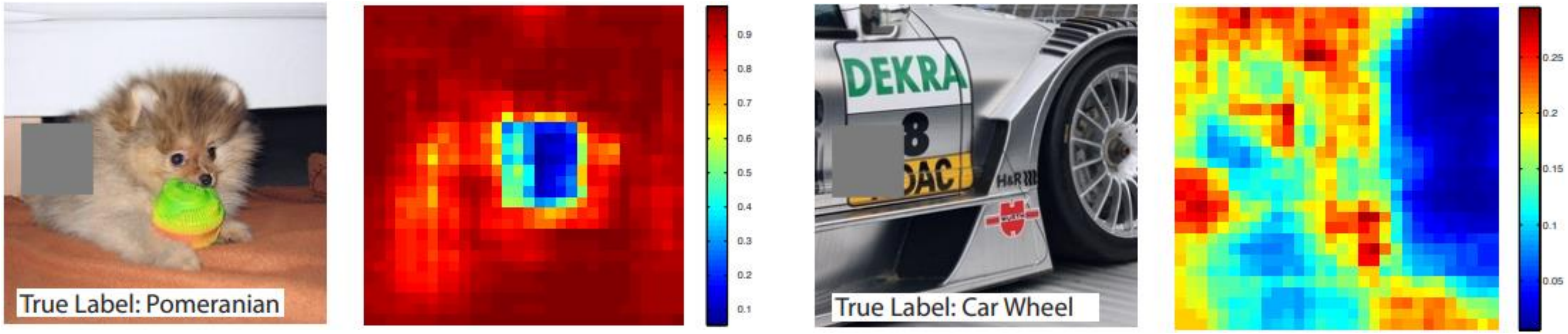
Slides and Resources

- What is a local *post hoc* explanation?
  - Some description of the model's local decision making logic
- If this training instance is not present in the dataset, the model will make a different prediction.



Feature Attributions (Saliency Maps)

Slides and Resources



Occlusion Saliency (Zeiler and Fergus, 2014)

Feature Attributions (Saliency Maps)

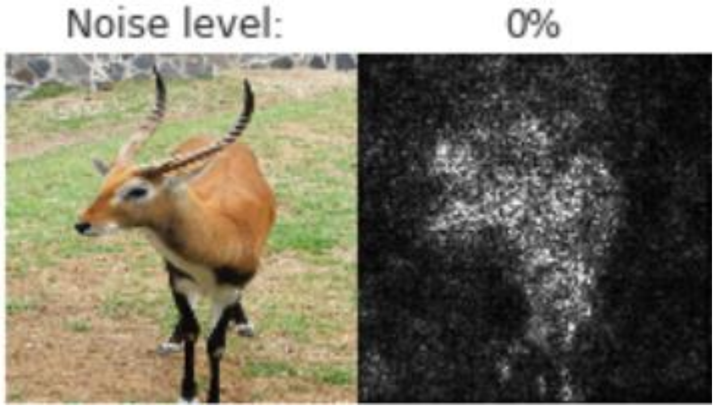
Slides and Resources

- Vanilla gradient: $s = \nabla_x f(x)$

Feature Attributions (Saliency Maps)

Slides and Resources

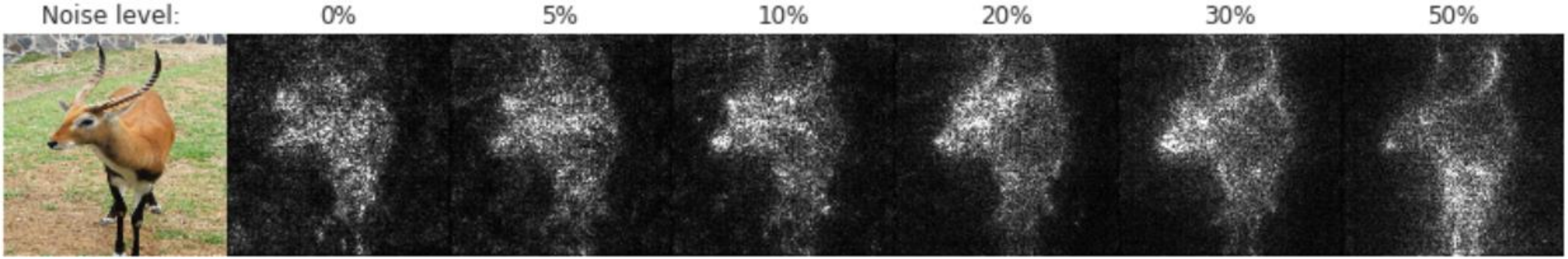
- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x + \epsilon)]$; ϵ_i independent for every x_i



Feature Attributions (Saliency Maps)

Slides and Resources

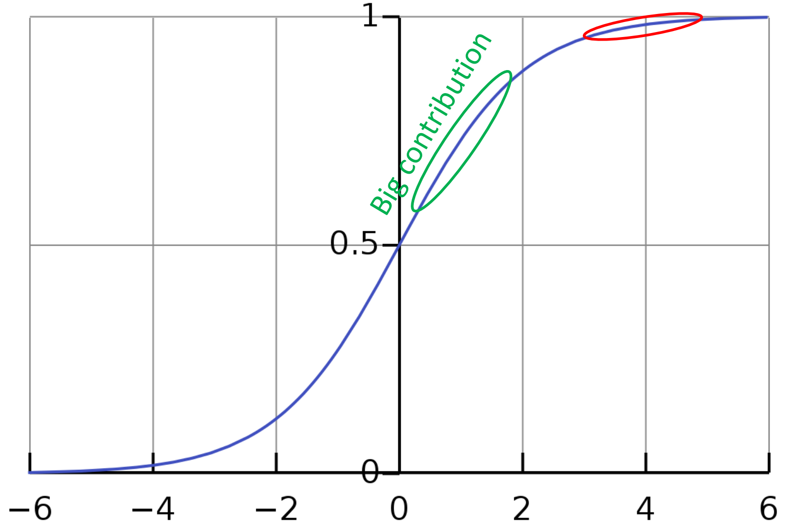
- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x + \epsilon)]$; ϵ_i independent for every x_i



Feature Attributions (Saliency Maps)

Slides and Resources

- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x+\epsilon)]$; ϵ_i independent for every x_i
- Integrated gradient: $s = \int_0^1 \nabla_x f\big(x_0 + u(x - x_0)\big)\,\mathrm{d}u$



Feature Attributions (Saliency Maps)

Slides and Resources

- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x+\epsilon)];\ \epsilon_i$ independent for every x_i
- Integrated gradient: $s = \int_0^1 \nabla_x f\big(x_0 + u(x - x_0)\big)\,\mathrm{d}u$
- Occlusion: $s_i = f(x) - f(x_{-i})$

Feature Attributions (Saliency Maps)

Slides and Resources

- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x + \epsilon)]$; ϵ_i independent for every x_i
- Integrated gradient: $s = \int_0^1 \nabla_x f\big(x_0 + u(x - x_0)\big)\, \mathrm{d}u$
- Occlusion: $s_i = f(x) - f(x_{-i})$
- LIME: $f \sim s^T m + b$

Feature mask m	Masked sentence	$f(\cdot)$
0, 0, 0, 0	“”	0.49
0, 0, 0, 1	“beautiful”	0.9
0, 0, 1, 0	“and”	0.52
0, 0, 1, 1	“and beautiful”	0.91
…		
1, 1, 1, 0	“It’s good and”	0.89
1, 1, 1, 1	“It’s good and beautiful”	0.95

Feature Attributions (Saliency Maps)

Slides and Resources

- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x + \epsilon)];\ \epsilon_i$ independent for every x_i
- Integrated gradient: $s = \int_0^1 \nabla_x f\big(x_0 + u(x - x_0)\big)\,\mathrm{d}u$
- Occlusion: $s_i = f(x) - f(x_{-i})$
- LIME: $f \sim s^T m + b$
- SHAP: $s_i = \sum_{S \subseteq F \setminus \{i\}} \frac{(|S|!(|F|-|S|-1)!)}{|F|!}\left[f_{S \cup \{i\}}\big(x_{S \cup \{i\}}\big) - f_S(x_S)\right]$

Feature Attributions (Saliency Maps)

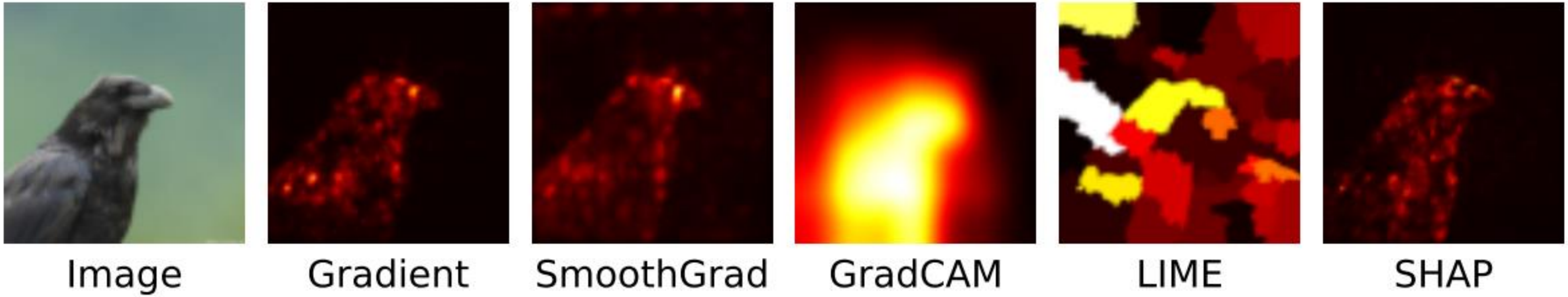
Slides and Resources

- Vanilla gradient: $s = \nabla_x f(x)$
- SmoothGrad: $s = \mathbb{E}_\epsilon[\nabla_x f(x+\epsilon)]$; ϵ_i independent for every x_i
- Integrated gradient: $s = \int_0^1 \nabla_x f\big(x_0 + u(x - x_0)\big)\,\mathrm{d}u$
- Occlusion: $s_i = f(x) - f(x_{-i})$
- LIME: $f \sim s^T m + b$
- SHAP: $s_i = \sum_{S \subseteq F \setminus \{i\}} \frac{(|S|!(|F|-|S|-1)!)}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right]$

$$e = D(x) \in \mathbb{R}^L$$

Feature Attributions (Saliency Maps)

Slides and Resources



$$e = D(x) \in \mathbb{R}^L$$

Evaluating Explanations

Slides and Resources

Problem: don't know how models work

Evaluating Explanations

Slides and Resources

Problem: don't know how models work → Solution: develop explanation methods

Evaluating Explanations

Slides and Resources

Problem: don't know how models work

Solution: develop explanation methods

Evaluation: compare with ground truth working mechanism

Evaluating Explanations

Problem: don't know how models work

Solution: develop explanation methods

Evaluation: compare with ground truth working mechanism

Evaluating Explanations

Slides and Resources

Problem: don't know how models work

~~Solution: develop explanation methods~~

Evaluation: compare with ground truth working mechanism

Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal

Proxy Metrics for Explanation Quality

- Feature importance $\Leftrightarrow$ model prediction change with feature removal

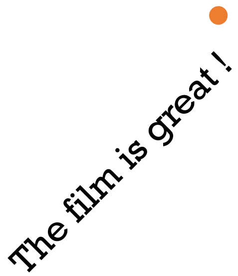
Gradient

Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal

Gradient



Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal

Gradient

Slides and Resources

Proxy Metrics for Explanation Quality

- Feature importance ⟺ model prediction change with feature removal

Gradient

Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal

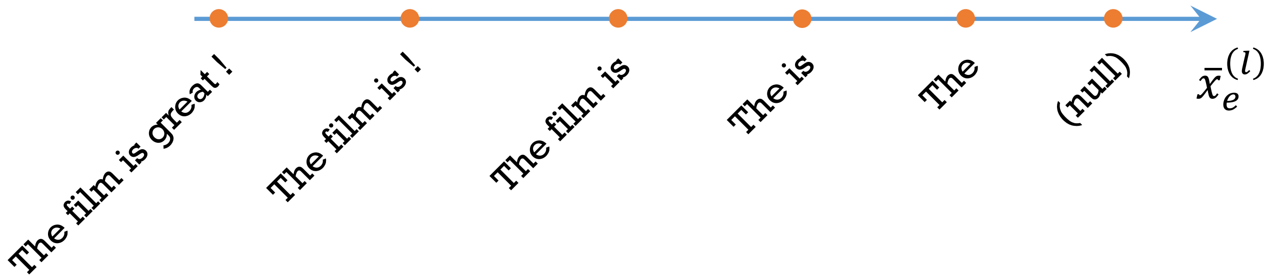
Gradient

Proxy Metrics for Explanation Quality

Slides and Resources

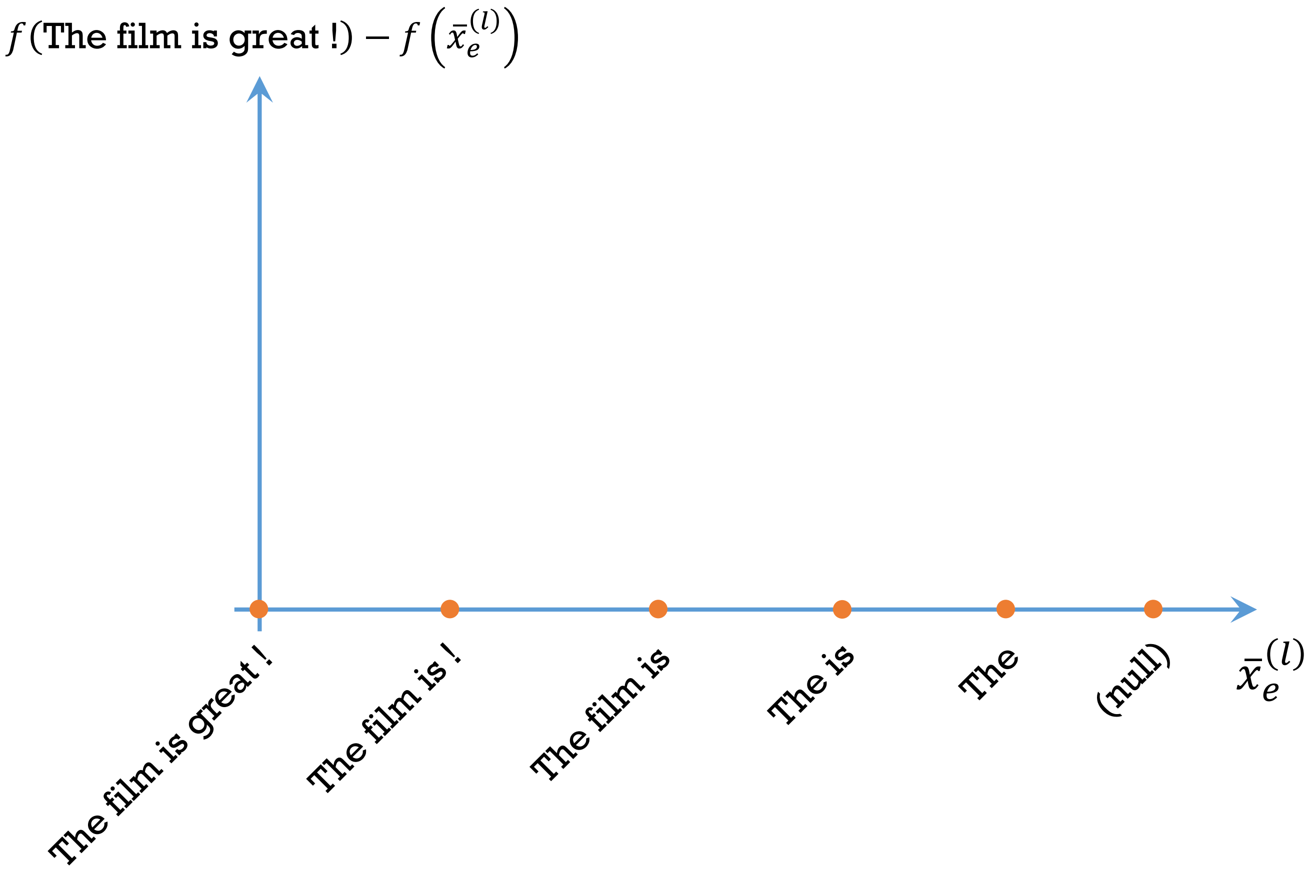
- Feature importance ⟺ model prediction change with feature removal

Gradient



Proxy Metrics for Explanation Quality

- Feature importance $\Leftrightarrow$ model prediction change with feature removal

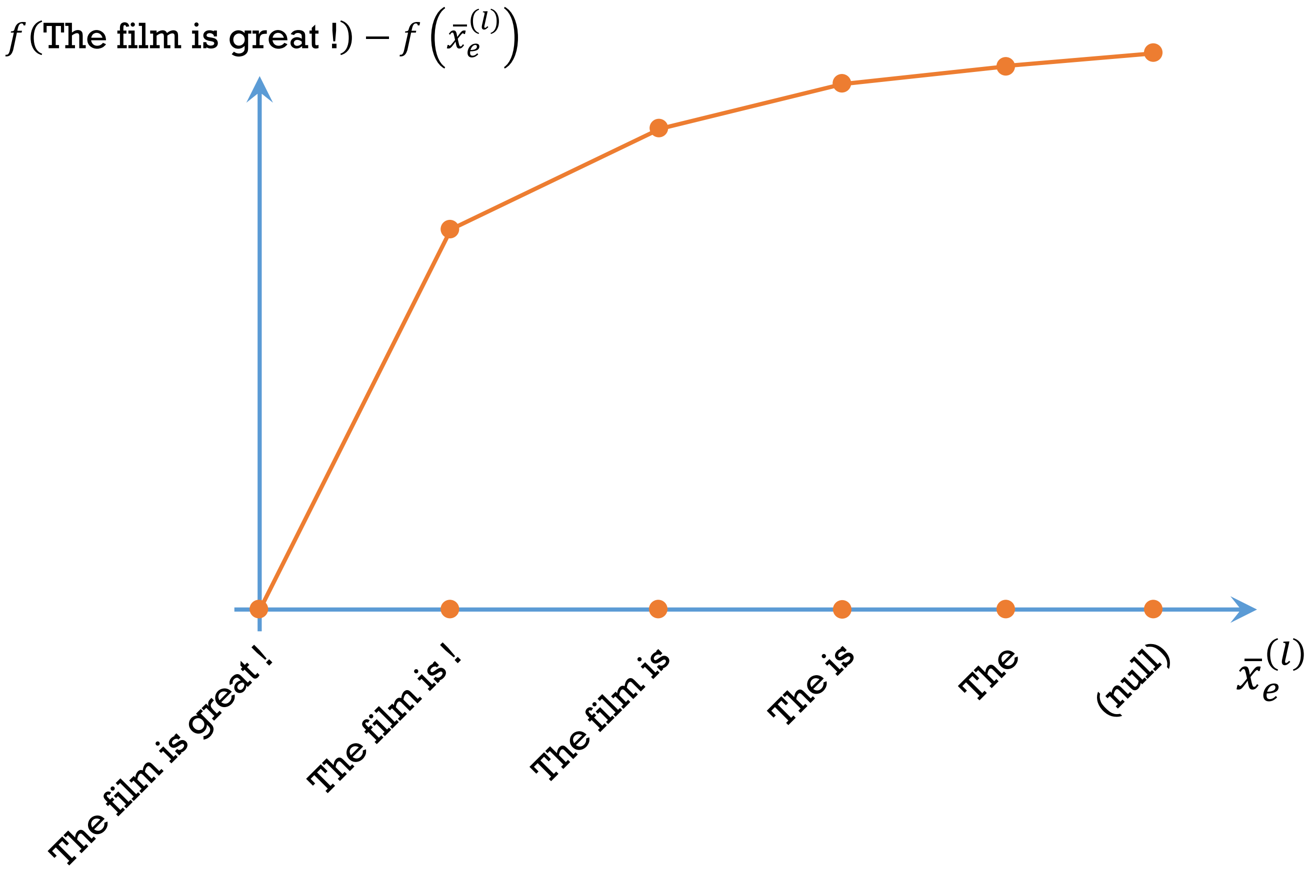


Gradient

Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal

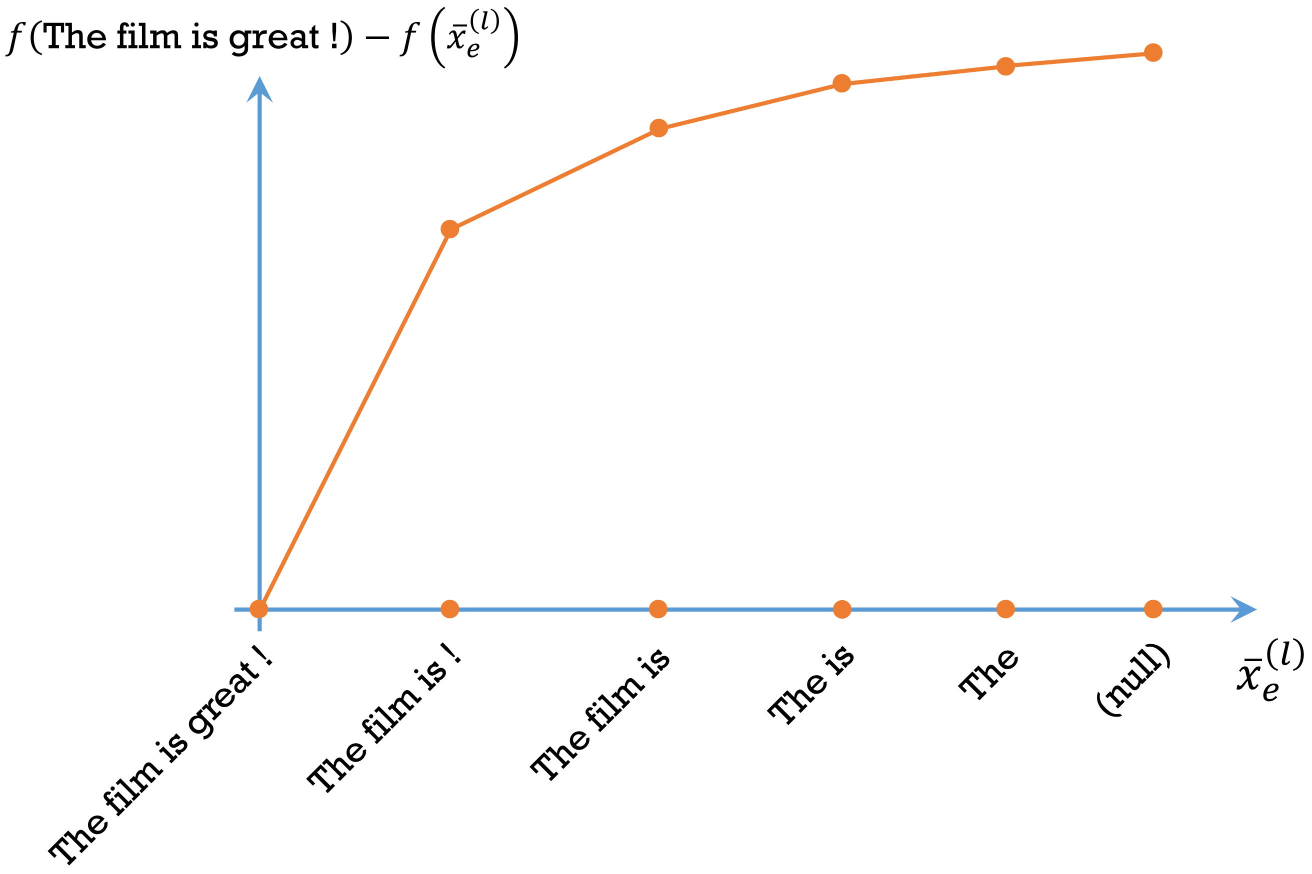


Gradient

Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal



Gradient

Comprehensiveness

$$\kappa(x,e) = \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

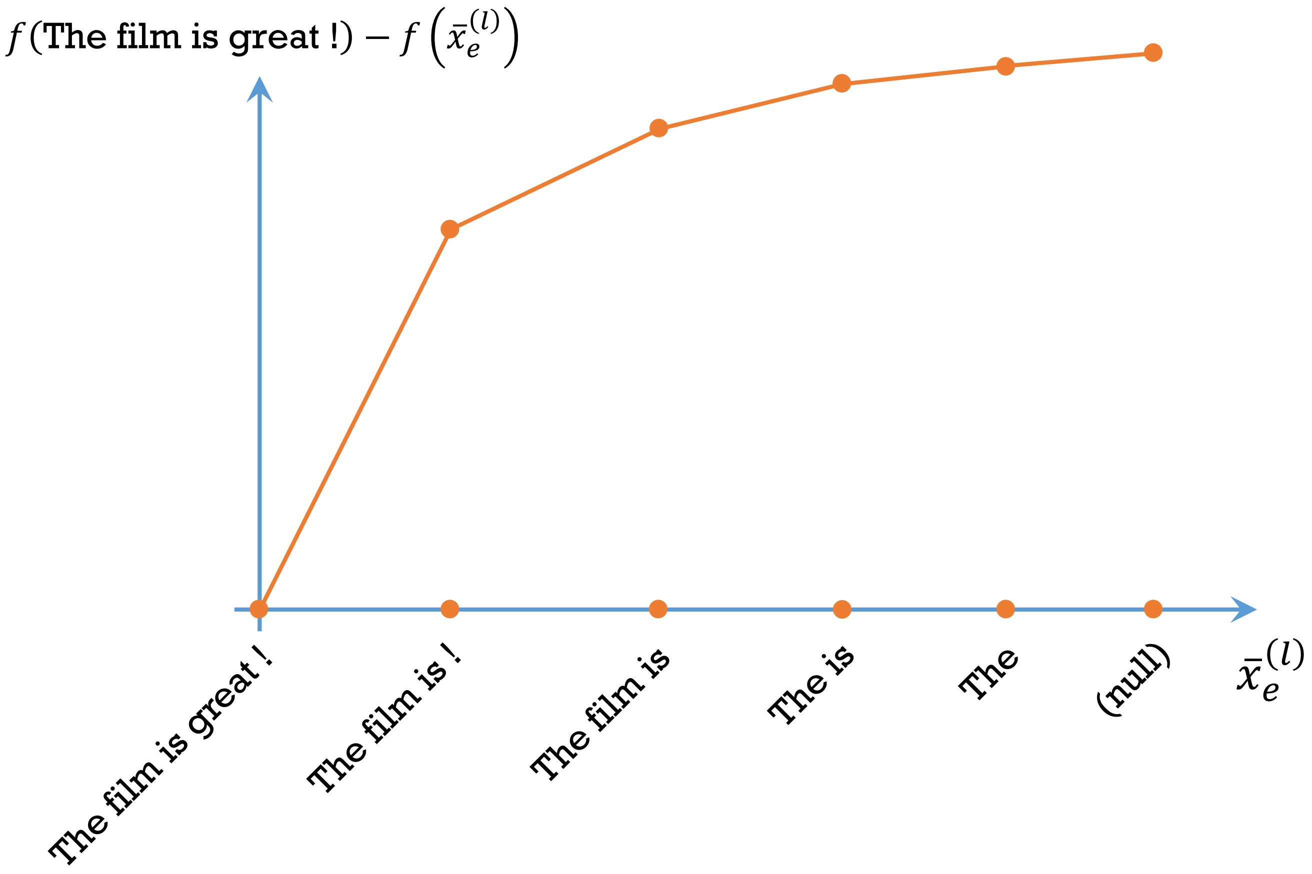
$\bar{x}^{(l)}$: input x with l most important features removed according to e

$e \in \mathbb{R}^D$ for D-dim input

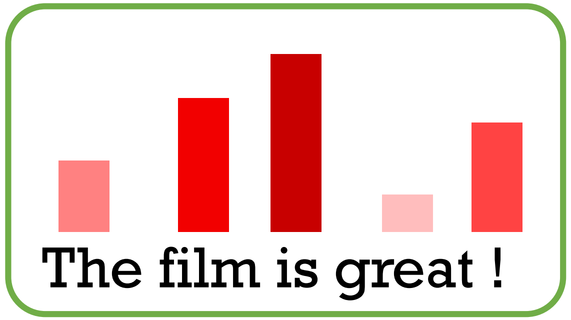
Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal



Gradient

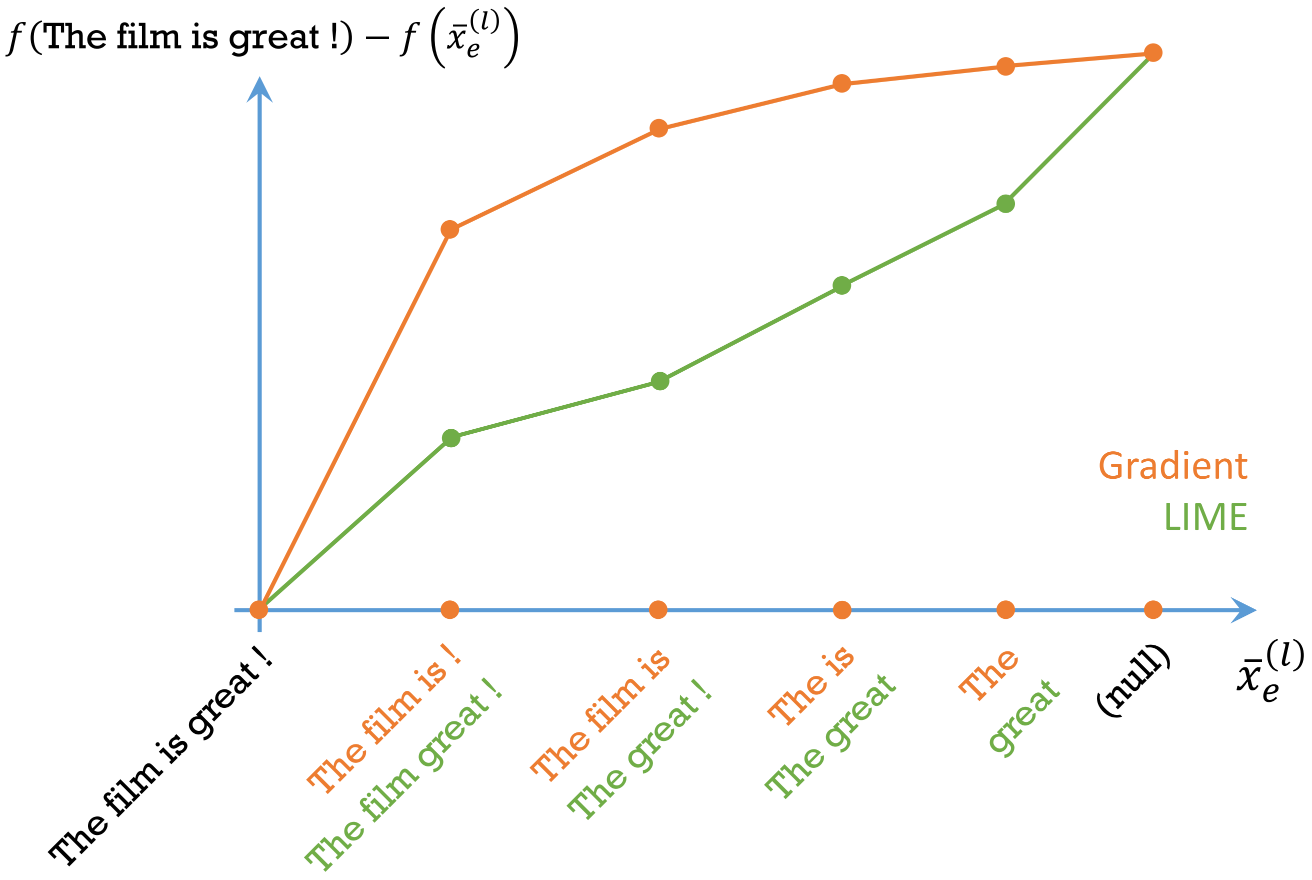


LIME

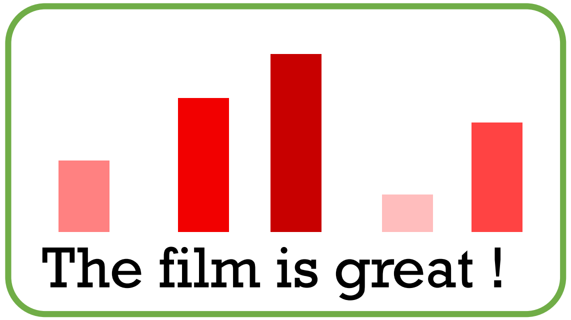
Proxy Metrics for Explanation Quality

Slides and Resources

- Feature importance ⟺ model prediction change with feature removal



Gradient



LIME

Common Evaluation Metrics

- Comprehensiveness and sufficiency
    - Also known as deletion and insertion metrics
    - Comprehensiveness also known as area over perturbation curve (AoPC)

Common Evaluation Metrics

Slides and Resources

- Comprehensiveness and sufficiency
  - Also known as deletion and insertion metrics
  - Comprehensiveness also known as area over perturbation curve (AoPC)
- Decision flip rate under most important feature removal
- Number of removals required for decision flip
- Prediction change rank correlation
- Etc.

Slides and Resources

Common Evaluation Metrics

- Comprehensiveness and sufficiency
  - Also known as deletion and insertion metrics
  - Comprehensiveness also known as area over perturbation curve (AoPC)
- Decision flip rate under most important feature removal
- Number of removals required for decision flip
- Prediction change rank correlation
- Etc.

Definition: $e = D(x) \in \mathbb{R}^D$

Evaluation: $q = E(x, e) \in \mathbb{R}$

Slides and Resources

Evaluating Explanations

Problem: don't know how models work

Solution: develop explanation methods

~~Evaluation: compare with ground truth working mechanism~~

Define an alternative notion of explanation quality

Evaluating Explanations

Slides and Resources

Problem: don't know how models work

Solution: develop explanation methods

~~Evaluation: compare with ground truth working mechanism~~

Induce a specific ground truth working mechanism

Are Explanations (Necessarily) Correct?

Slides and Resources

If we know that a specific feature is crucial to the model prediction, can feature attribution explanations identify its importance?

Dataset Modification

Slides and Resources

X

Y


→ Crow


→ Crow


→ Sparrow

X: original input image

Y: original output label

Dataset Modification

Slides and Resources

X

$Y \dashrightarrow \hat{Y} \sim \mathrm{Ber}(0.5)$

 → Crow

 Crow

 → Crow

 Sparrow

 → Sparrow

 Crow

X: original input image

Y: original output label

$\hat{Y}$: modified output label

Dataset Modification

Slides and Resources

$$X \longrightarrow \hat{X}$$

$$Y \dashrightarrow \hat{Y} \sim \mathrm{Ber}(0.5), \quad \hat{Y} \rightarrow \hat{X}$$

 → Crow | 

← Crow

 → Crow | 

← Sparrow

 → Sparrow | 

← Crow

X: original input image

Y: original output label

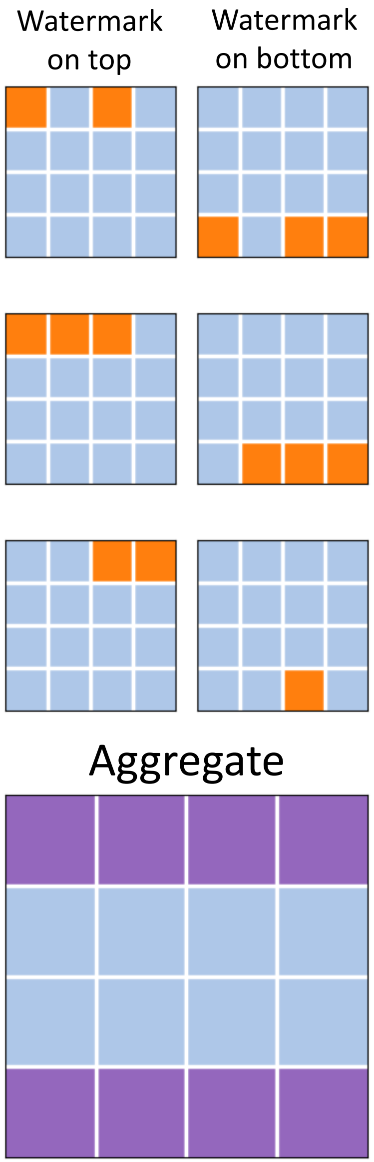
$\hat{Y}$: modified output label

$\hat{X}$: modified input image

Dataset Modification

$$X \longrightarrow \hat{X}$$

$$Y \dashrightarrow \hat{Y} \sim \mathrm{Ber}(0.5)$$



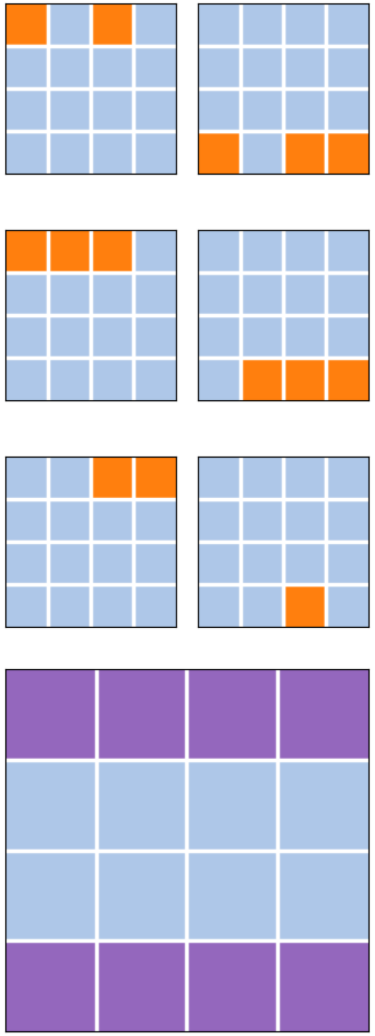
Slides and Resources

Dataset Modification

Slides and Resources

Slides and
Resources

Dataset Modification

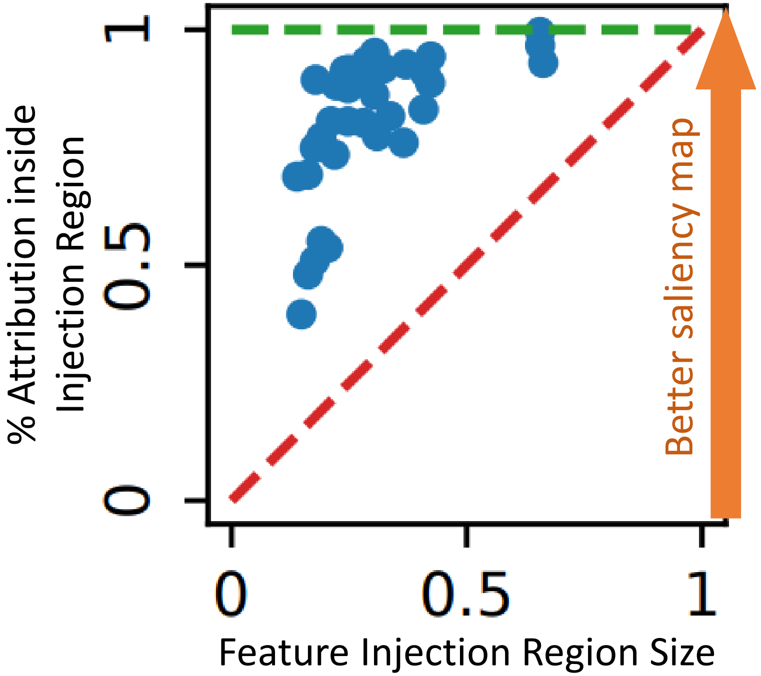


$$\%\ \text{Attribution} = \frac{\sum A}{\sum A + \sum A}$$

Evaluating Image Saliency Maps

Slides and Resources

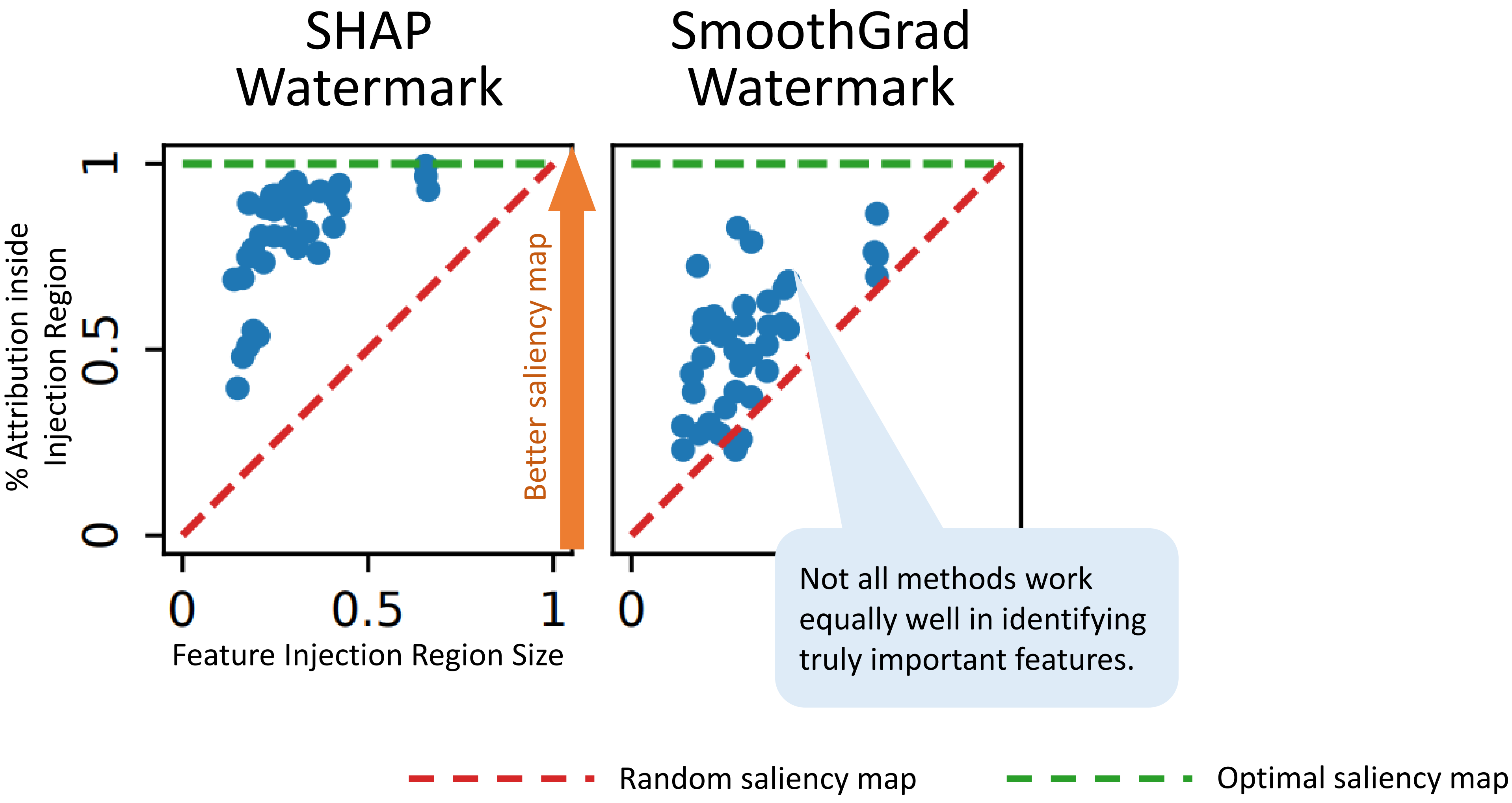
SHAP
Watermark



Random saliency map
Optimal saliency map

Evaluating Image Saliency Maps

Slides and Resources

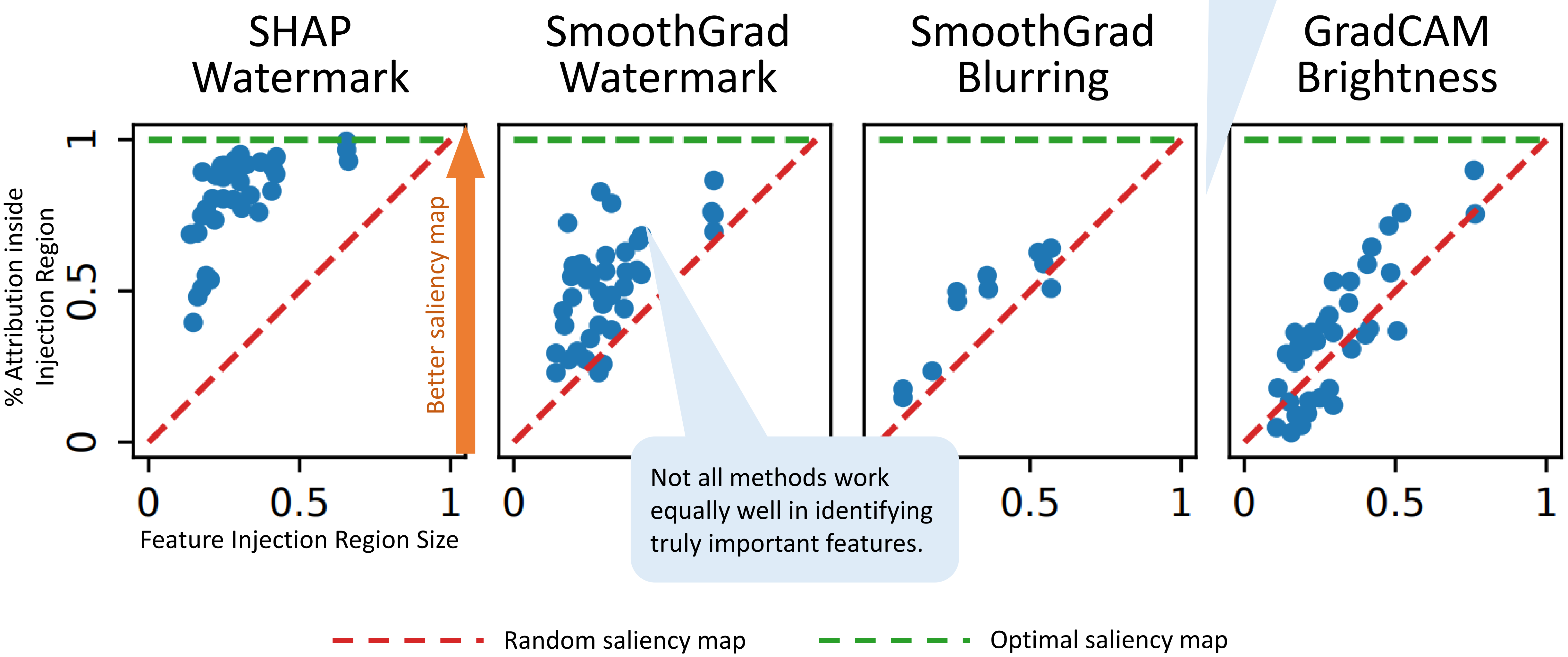


Not all methods work equally well in identifying truly important features.

Evaluating Image Saliency Maps

Some methods even struggle against the random baseline on certain feature types.

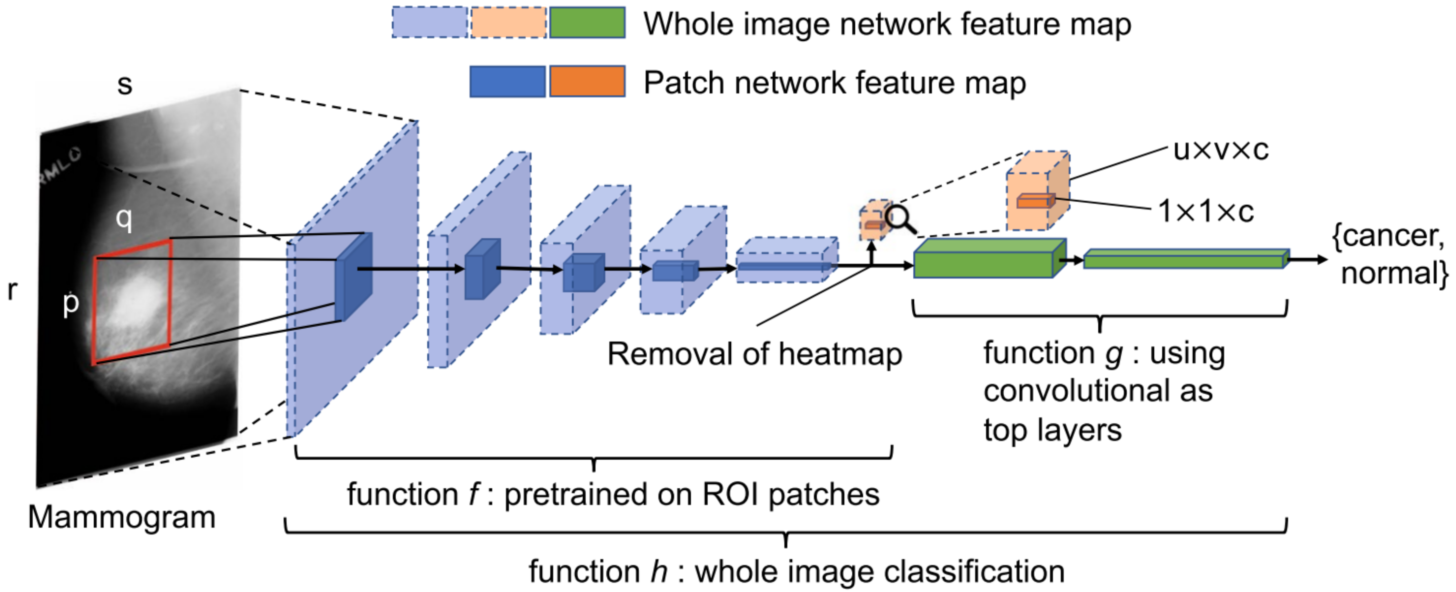
Slides and Resources



Not all methods work equally well in identifying truly important features.

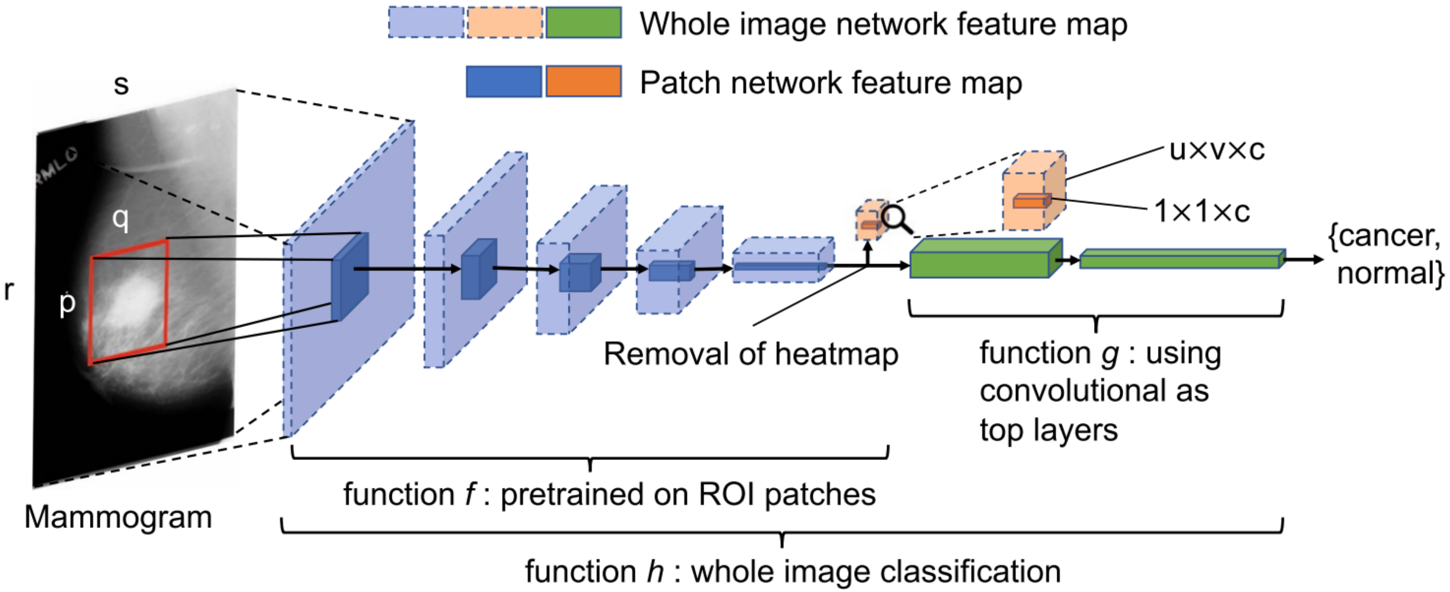
Practical Implications

Slides and Resources



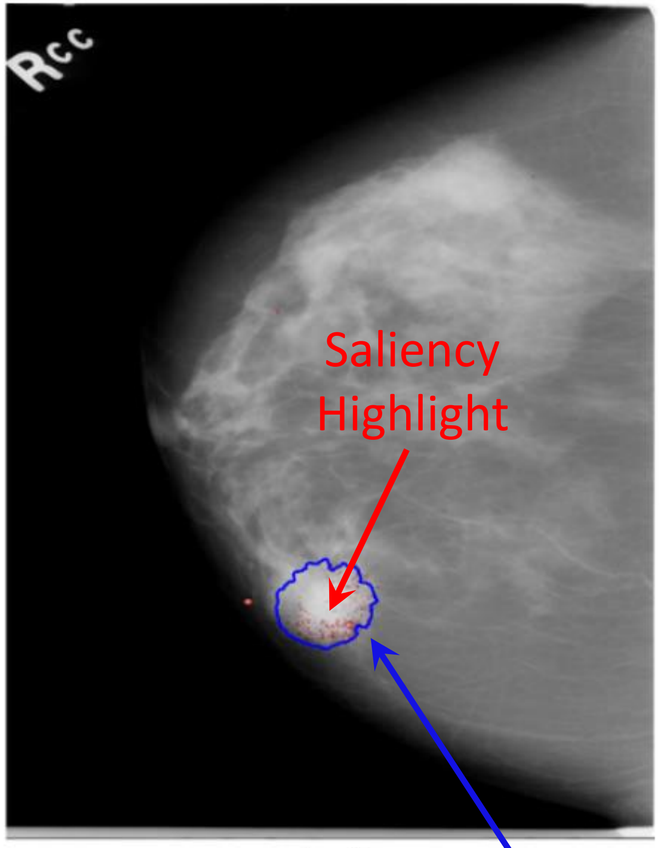
Practical Implications

Slides and Resources



“A saliency map illustrates which area of the input image is considered to be responsible for the cancer […] Figure 4a […] shows that the image classifier was able to correctly locate the cancerous region on which its decision was based.”

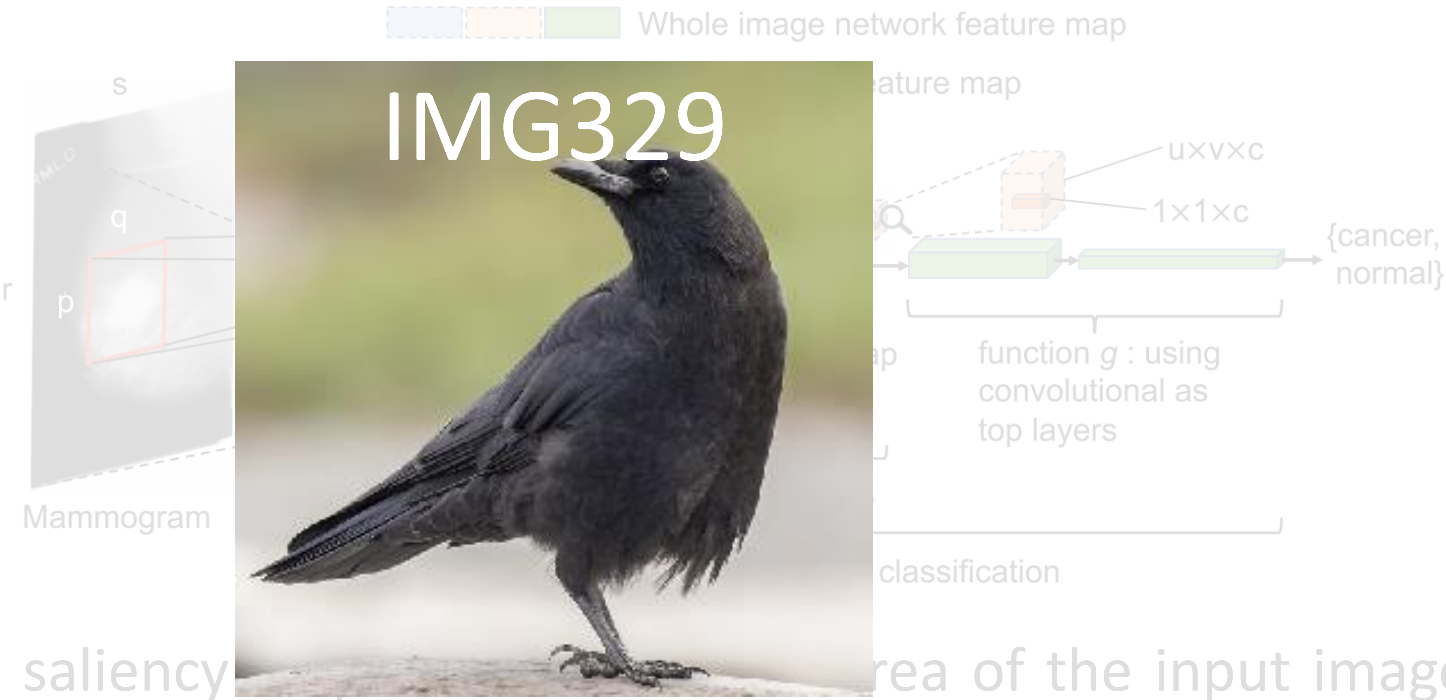
(Shen et al., *Scientific Reports*, 2019)



(a)

Practical Implications

Slides and Resources

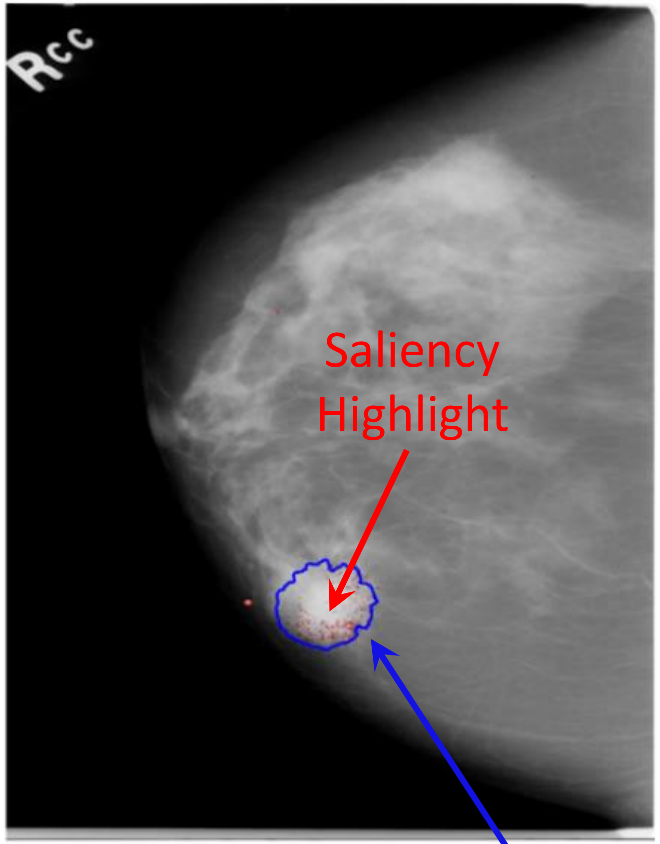


Known spurious correlation

Some explainers don't work

"A saliency rea of the input image is considered to he cancer [...] Figure 4a [...] shows that the image classifier was able to correctly locate the cancerous region on which its decision was based."

(Shen et al., *Scientific Reports*, 2019)

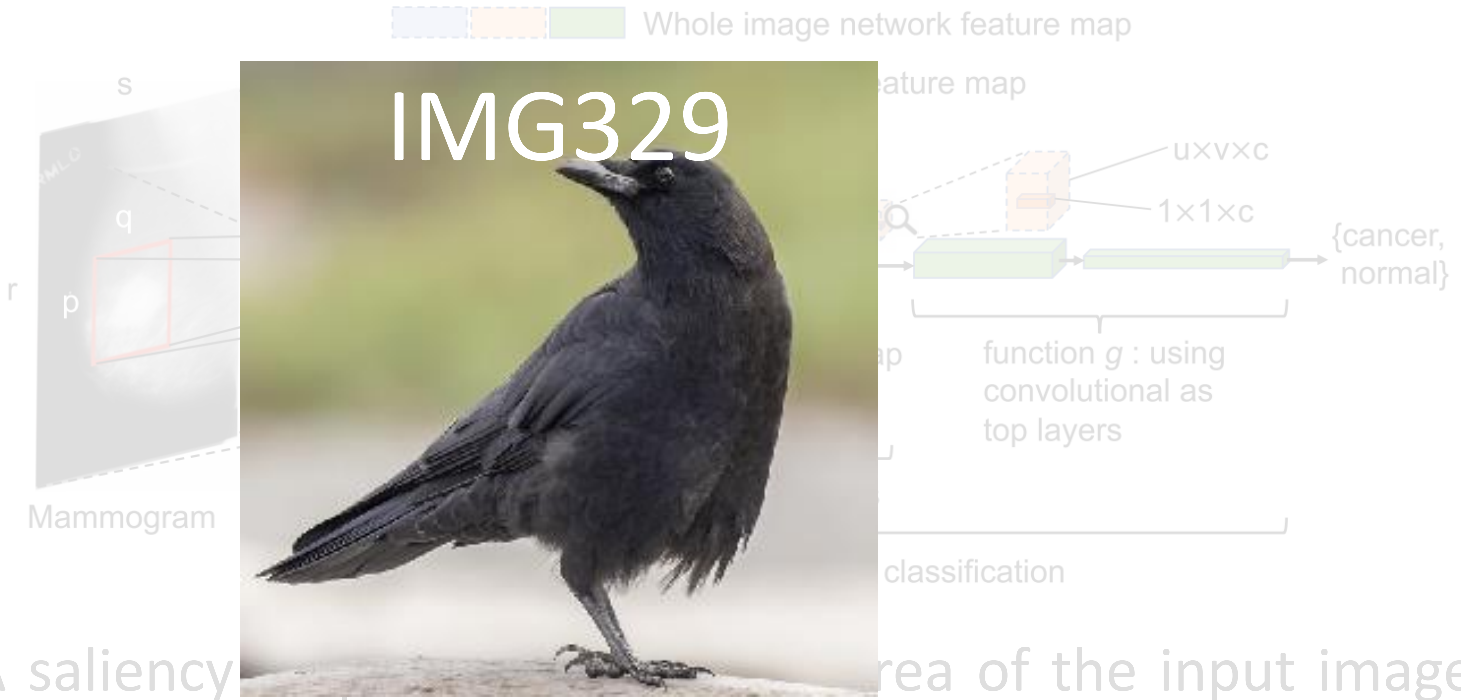


(a)

Tumor Segmentation

Practical Implications

Slides and Resources

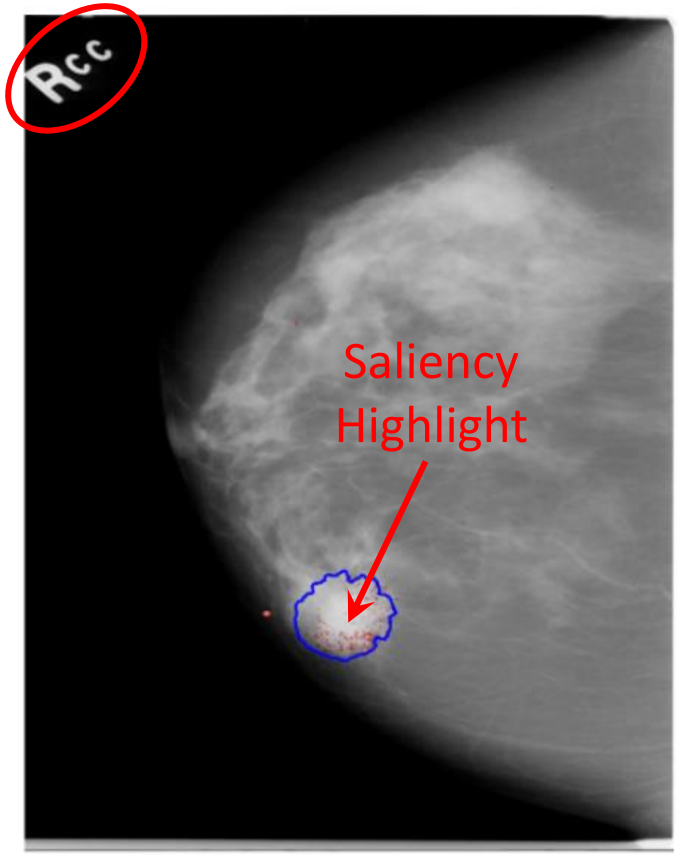


Known spurious correlation

Some explainers don't work

“A saliency ... rea of the input image is considered to ... he cancer [...] Figure 4a [...] shows that the image classifier was able to correctly locate the cancerous region on which its decision was based.”

(Shen et al., *Scientific Reports*, 2019)



Unknown spurious correlation

Can we trust these explainers?

Explanation Understandability

Slides and Resources

Do the explanations correctly explain the model prediction logic?

Explanation Understandability

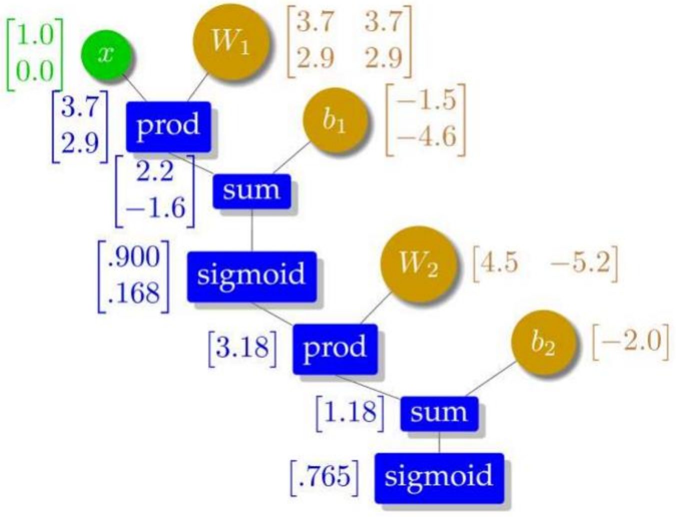
Slides and Resources

Do the explanations correctly explain the model prediction logic?

Do people correctly understand the model prediction logic from the explanations?

Correct but Not Understandable Explanations

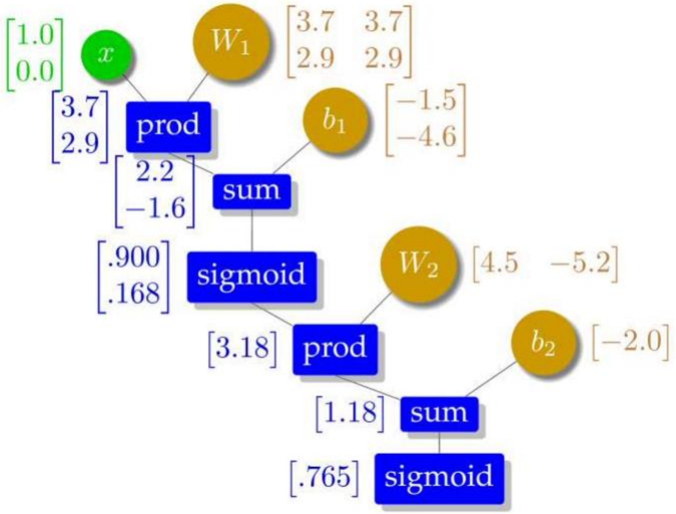
Slides and Resources

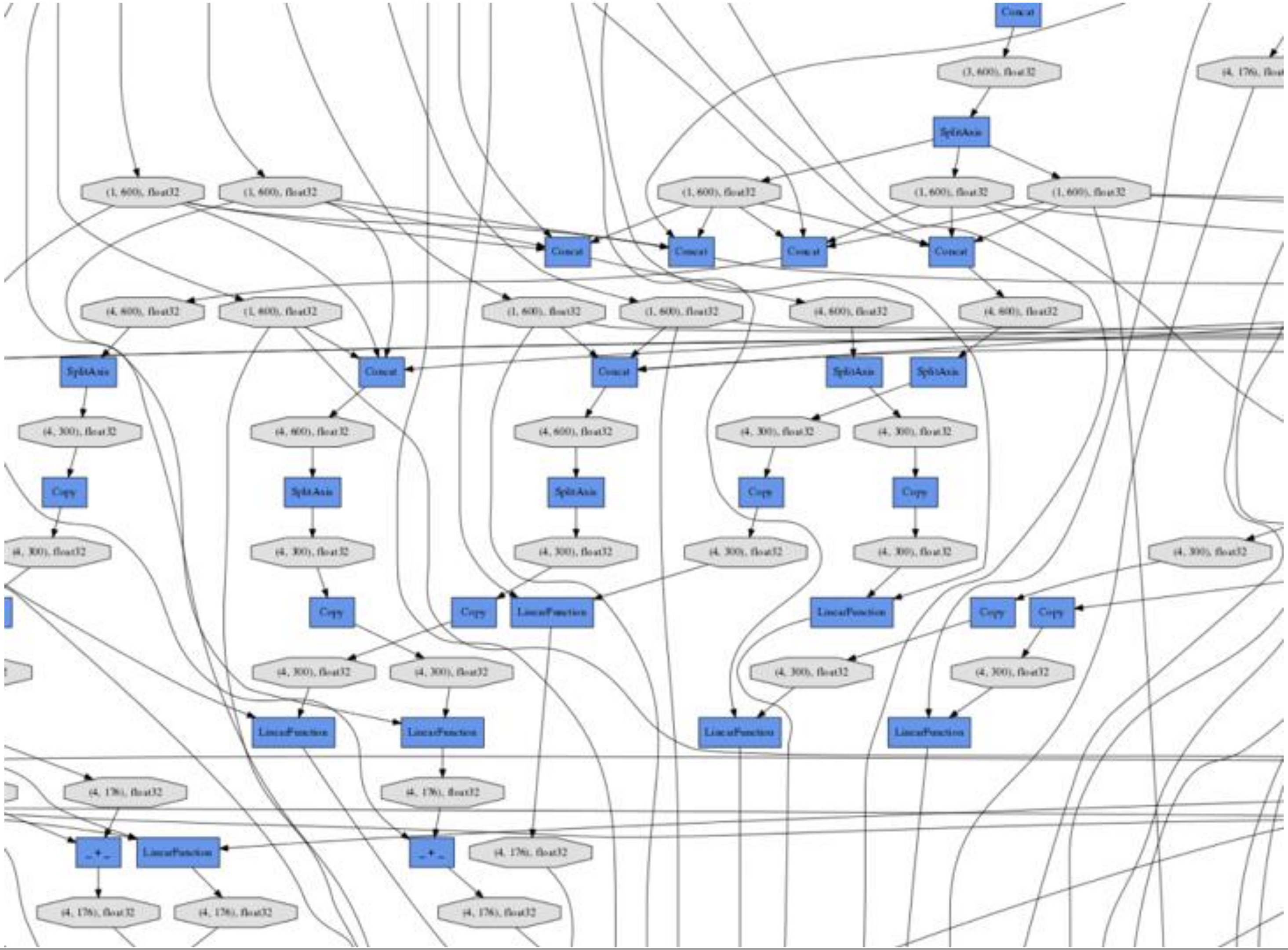


Computation trace: a fully correct explanation for the prediction

Correct but Not Understandable Explanations

Slides and Resources





Computation trace: a fully correct
but totally not understandable
explanation for the prediction

Slides and Resources

Understanding a Sentiment Classifier

Data Instance

Input: As shaky as the plot is , Kaufman 's script is still memorable for some great one-liners .
Label: Positive
Model Prediction: Positive

Local Explanation

SHAP score

memorable:	0.48
great:	0.37
for:	-0.02
one-liners:	-0.14
shaky:	-0.39

Understanding a Sentiment Classifier

Slides and Resources

Data Instance

Input: As shaky as the plot is , Kaufman 's script is still memorable for some great one-liners .
Label: Positive
Model Prediction: Positive

Local Explanation

SHAP score

memorable:	0.48
great:	0.37
for:	-0.02
one-liners:	-0.14
shaky:	-0.39

General Rule

A positive sentiment word contributes very positively to the prediction.

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable:  0.48
great:      0.37
```

Label = Positive
Prediction = Positive

A positive sentiment word contributes very positively to the prediction

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable: 0.48
great:     0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive: 0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word contributes very positively to the prediction

Supporting instance

Opposing instance

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable: 0.48
great:     0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive: 0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word contributes very positively to the prediction…

unless it is negated

Supporting instance

Opposing instance

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable:  0.48
great:      0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive:  0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word contributes very positively to the prediction…

unless it is negated

Daring, **mesmerizing** and exceed-ingly hard to forget.

```
Daring:         0.13
mesmerizing:    0.06
```

Label = Positive
Prediction = Positive

Supporting instance

Opposing instance

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable:  0.48
great:      0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive:  0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word contributes very positively to the prediction…

unless it is negated, or near another positive word

Daring, **mesmerizing** and exceed-ingly hard to forget.

```
Daring:         0.13
mesmerizing:    0.06
```

Label = Positive
Prediction = Positive

Supporting instance

Opposing instance

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable: 0.48
great:     0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive: 0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word contributes very positively to the prediction…

unless it is negated, or near another positive word

Daring, **mesmerizing** and exceed-ingly hard to forget.

```
Daring:       0.13
mesmerizing:  0.06
```

Label = Positive
Prediction = Positive

Ranks among Willams' **best** screen work.

```
best: 0.05
```

Label = Positive
Prediction = Positive

Supporting instance

Opposing instance

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable:  0.48
great:      0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive:  0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word sometimes contributes very positively to the prediction…

unless it is negated, or near another positive word

Daring, **mesmerizing** and exceed-ingly hard to forget.

```
Daring:         0.13
mesmerizing:    0.06
```

Label = Positive
Prediction = Positive

Ranks among Willams' **best** screen work.

```
best:  0.05
```

Label = Positive
Prediction = Positive

Supporting instance

Opposing instance

Understanding a Sentiment Classifier

Slides and Resources

As shaky as the plot is, Kaufman's script is still **memorable** for some **great** one-liners.

```
memorable:  0.48
great:      0.37
```

Label = Positive
Prediction = Positive

Just because it really happened to you, honey, does n't mean that it's **attractive** to anyone else.

```
attractive:  0.00
```

Label = Negative
Prediction = Negative

A positive sentiment word sometimes contributes very positively to the prediction…

unless it is negated, or near another positive word

Daring, **mesmerizing** and exceed-ingly hard to forget.

```
Daring:         0.13
mesmerizing:    0.06
```

Label = Positive
Prediction = Positive

Ranks among Willams' **best** screen work.

```
best:  0.05
```

Label = Positive
Prediction = Positive

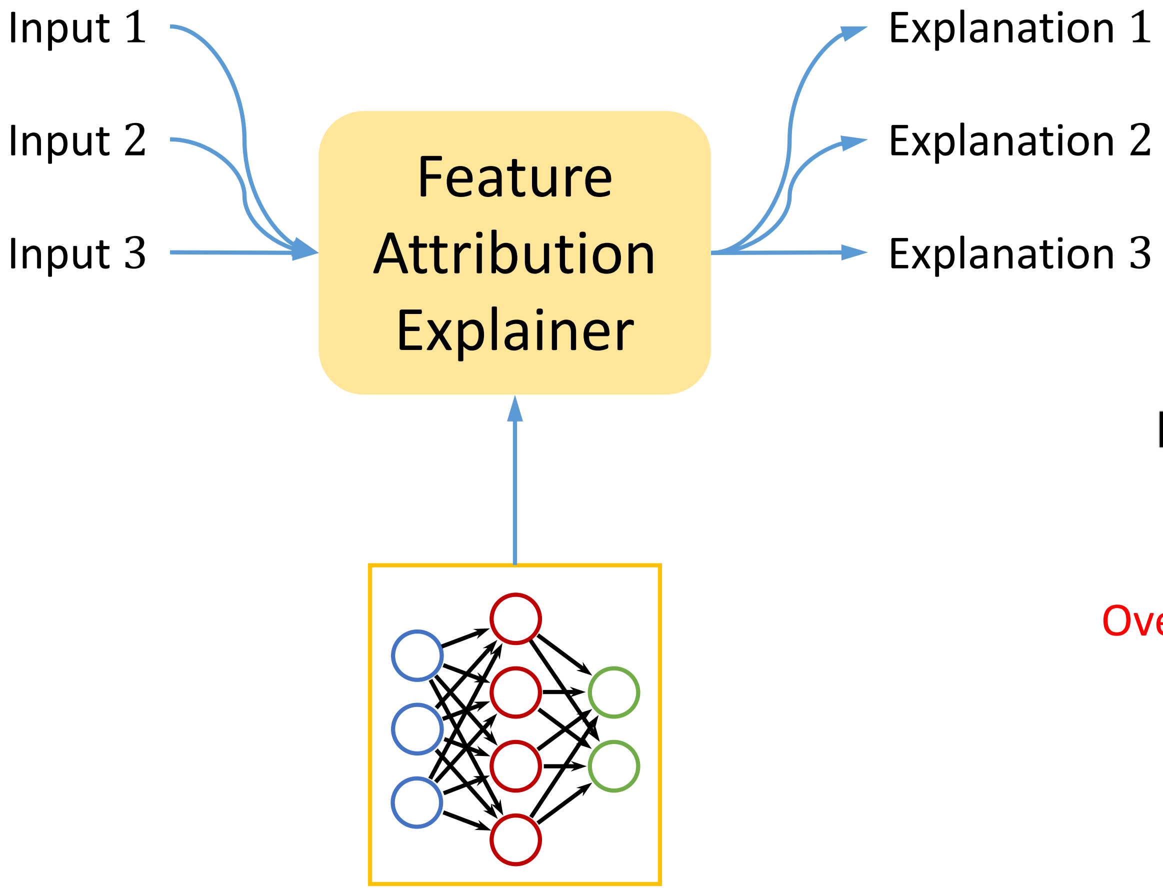
(many more counter-examples not shown)

Supporting instance

Opposing instance

Explanation Summary (ExSum)

Slides and Resources



Over-generalization

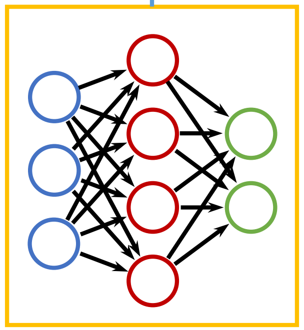
Explanation Summary (ExSum)

Slides and Resources

Input 1
Input 2
Input 3
…
Input N

Feature Attribution Explainer

Explanation 1
Explanation 2
Explanation 3
…
Explanation N



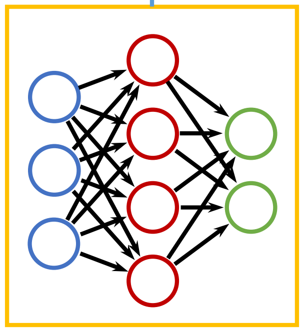
Heavy cognitive workload

Explanation Summary (ExSum)

Slides and Resources

Input 1
Input 2
Input 3
…
Input N

Feature Attribution Explainer



Explanation 1
Explanation 2
Explanation 3
…
Explanation N

ExSum

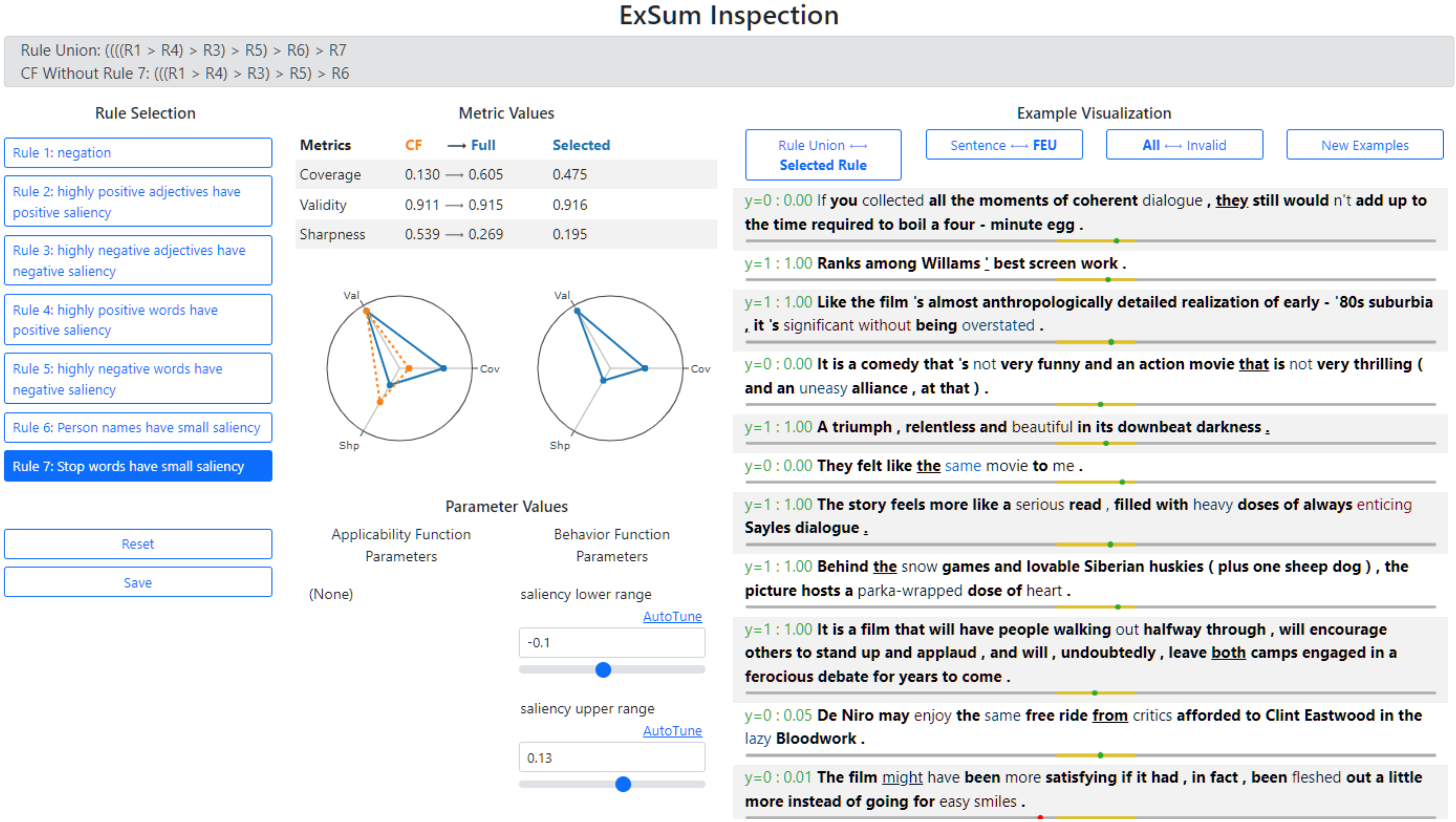
Metric Values:
- Coverage
- Validity
- Sharpness

Model Understanding 1
Model Understanding 2
…
Model Understanding M
$(M \ll N)$

Systematic and quantitative model understanding

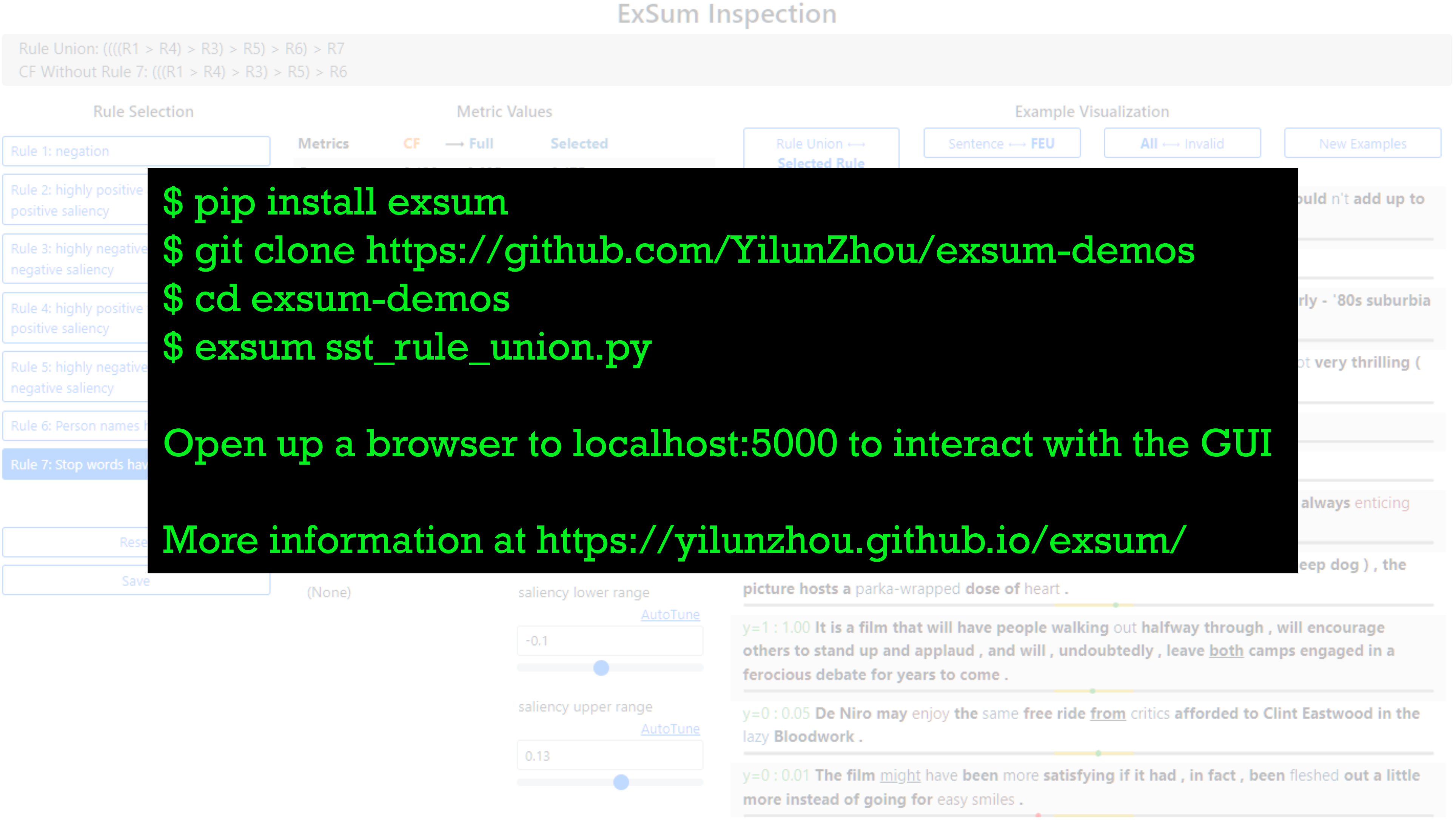
GUI for Developing ExSum Rules

Slides and Resources



GUI for Developing ExSum Rules

Slides and Resources

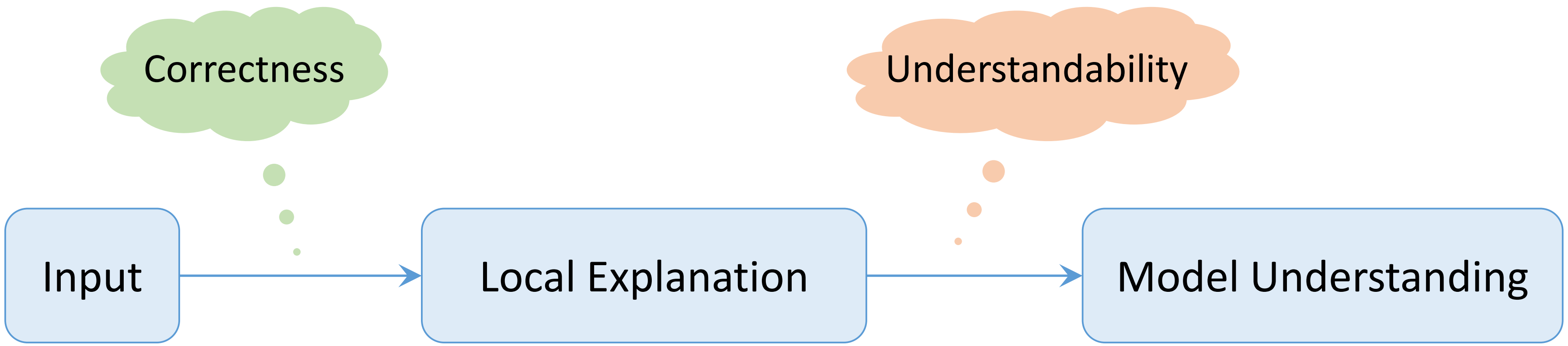



```
$ pip install exsum
$ git clone https://github.com/YilunZhou/exsum-demos
$ cd exsum-demos
$ exsum sst_rule_union.py
```

Open up a browser to localhost:5000 to interact with the GUI

More information at https://yilunzhou.github.io/exsum/

The Many Faces of Understandability



Slides and Resources

The Many Faces of Understandability

Correctness

Understandability

Input → Local Explanation → Model Understanding

Human Alignment:
$e(x) \approx h(x)$ for human reasoning process $h(\cdot)$

Robustness:
$\left|\left|e(x_1) - e(x_2)\right|\right| \leq \gamma\left|\left|x_1 - x_2\right|\right|$

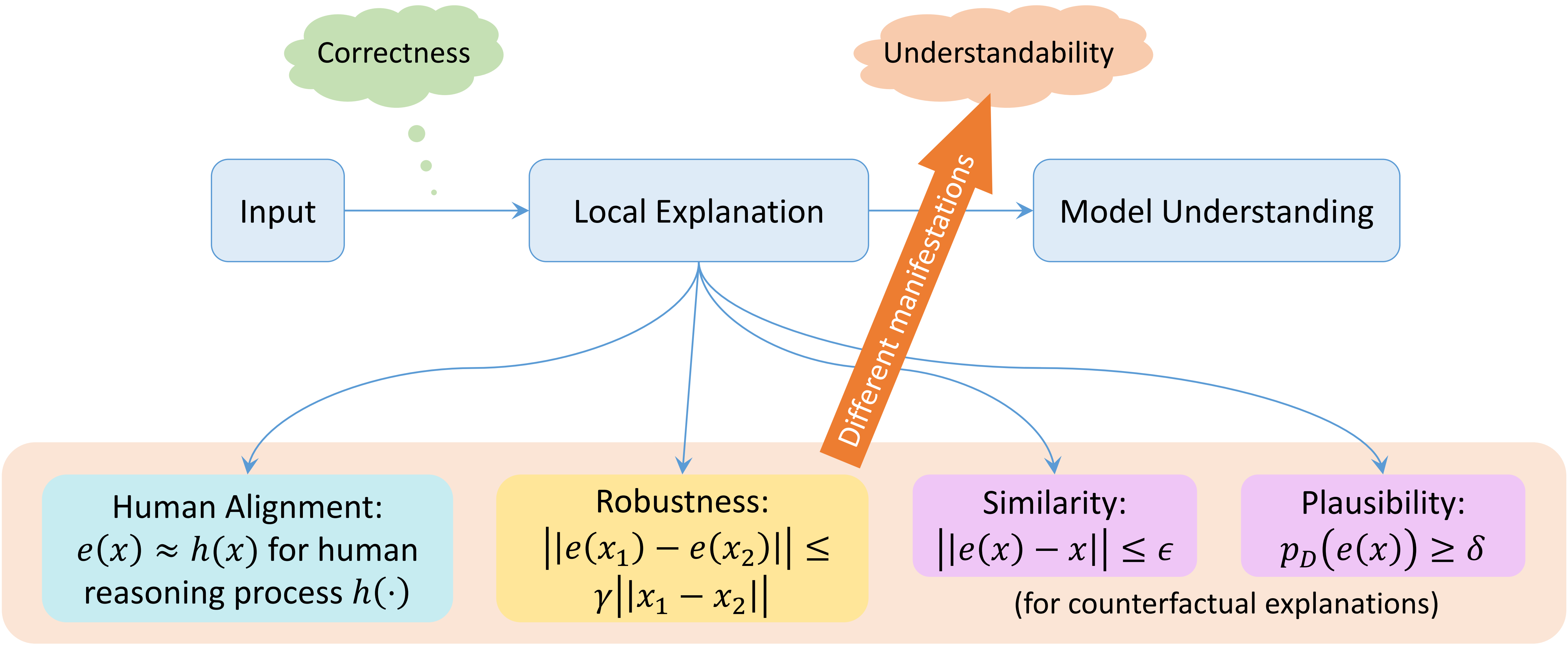
Similarity:
$\left|\left|e(x) - x\right|\right| \leq \epsilon$

Plausibility:
$p_D\big(e(x)\big) \geq \delta$

(for counterfactual explanations)

The Many Faces of Understandability

Slides and Resources



Definition vs. Evaluation

Slides and Resources

Definition

Gradient

$$D_g(x) = \nabla_x f(x)$$

Evaluation

Comprehensiveness

$$E_\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Definition vs. Evaluation

Slides and Resources

Definition

Gradient

$$D_g(x) = \nabla_x f(x)$$

Evaluation

Comprehensiveness

$$E_\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Gradient

$$E_g(x, e) = -\left|\left|\nabla_x f(x) - e\right|\right|$$

Definition vs. Evaluation

Slides and Resources

Definition

Gradient

$$D_g(x) = \nabla_x f(x)$$

Comprehensiveness

$$D_\kappa(x) = \operatorname*{argmax}_e \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Evaluation

Comprehensiveness

$$E_\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Gradient

$$E_g(x, e) = -\left|\left|\nabla_x f(x) - e\right|\right|$$

Definition-Evaluation Duality

Slides and Resources

Definition	Duality	Evaluation
Gradient $$D_g(x) = \nabla_x f(x)$$		Comprehensiveness $$E_\kappa(x, e) = \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$
Comprehensiveness $$D_\kappa(x) = \underset{e}{\operatorname{argmax}} \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$		Gradient $$E_g(x, e) = -\left\|\left|\nabla_x f(x) - e\right|\right\|$$

Slides and Resources

Definition-Evaluation Duality

Duality

Definition

Gradient

$$D_g(x) = \nabla_x f(x)$$

Comprehensiveness

$$D_\kappa(x) = \operatorname*{argmax}_e \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Evaluation

Comprehensiveness

$$E_\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Gradient

$$E_g(x, e) = -\left|\left|\nabla_x f(x) - e\right|\right|$$

Definition-Evaluation Duality

Slides and Resources

Definition — Duality — Evaluation

Definition

Gradient

$$D_g(x) = \nabla_x f(x)$$

$$D_g \gg^{\text{"ease"}} D_\kappa$$

Comprehensiveness

$$D_\kappa(x) = \underset{e}{\operatorname{argmax}} \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Evaluation

Comprehensiveness

$$E_\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

Gradient

$$E_g(x, e) = -\left|\left|\nabla_x f(x) - e\right|\right|$$

Definition-Evaluation Duality

Slides and Resources

Definition ⟷ Duality ⟷ Evaluation

Definition	Evaluation
Gradient $D_g(x) = \nabla_x f(x)$	Comprehensiveness $E_\kappa(x, e) = \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$
$D_g \gg$ "ease" D_κ **?**	
Comprehensiveness $D_\kappa(x) = \operatorname{argmax}_e \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$	Gradient $E_g(x, e) = -\lVert \nabla_x f(x) - e \rVert$

Beam Search Results

Slides and Resources

$$\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

$$e^* = \underset{e}{\operatorname{argmax}}\, \kappa(x, e)$$

Algorithm 1: Beam search for finding e^*.

Input: beam size B, metric m, sentence x of length L;
Let $e^{(0)}$ be an empty length-L explanation;
$\texttt{beams} \leftarrow \{e^{(0)}\}$;
for $l = 1, \ldots, L$ **do**
  $\texttt{beams} \leftarrow \bigcup_{e \in \texttt{beams}} \text{ext}(e, L - l + 1)$;
  $\texttt{beams} \leftarrow \text{choose\_best}(\texttt{beams}, B)$;
end
$e^* \leftarrow \text{choose\_best}(\texttt{beams}, 1)$;
$e^* \leftarrow \text{shift}(e^*)$;
return e^*;

Beam Search Results

Slides and Resources

$$\kappa(x,e) = \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

$$e^* = \underset{e}{\operatorname{argmax}}\ \kappa(x,e)$$

Algorithm 1: Beam search for finding e^*.

1 **Input**: beam size B, metric m, sentence x of length L;
2 Let $e^{(0)}$ be an empty length-L explanation;
3 $\texttt{beams} \leftarrow \{e^{(0)}\}$;
4 **for** $l = 1, \ldots, L$ **do**
5 $\quad \texttt{beams} \leftarrow \bigcup_{e \in \texttt{beams}} \mathrm{ext}(e, L - l + 1)$;
6 $\quad \texttt{beams} \leftarrow \mathrm{choose\_best}(\texttt{beams}, B)$;
7 **end**
8 $e^* \leftarrow \mathrm{choose\_best}(\texttt{beams}, 1)$;
9 $e^* \leftarrow \mathrm{shift}(e^*)$;
10 **return** e^*;

A worthy tribute to a great humanitarian and her vibrant ‘ co-stars . ’

So stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .

Beam Search Results

Slides and Resources

$$\kappa(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$

$$e^* = \operatorname*{argmax}_e \kappa(x, e)$$

Algorithm 1: Beam search for finding e^*.

Input: beam size B, metric m, sentence x of length L;
Let $e^{(0)}$ be an empty length-L explanation;
$\mathtt{beams} \leftarrow \{e^{(0)}\}$;
for $l = 1, \ldots, L$ **do**
  $\mathtt{beams} \leftarrow \bigcup_{e \in \mathtt{beams}} \mathrm{ext}(e, L-l+1)$;
  $\mathtt{beams} \leftarrow \mathrm{choose\_best}(\mathtt{beams}, B)$;
end
$e^* \leftarrow \mathrm{choose\_best}(\mathtt{beams}, 1)$;
$e^* \leftarrow \mathrm{shift}(e^*)$;
return e^*;

A worthy tribute to a great humanitarian and her vibrant ‘ co-stars . ’

So stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .

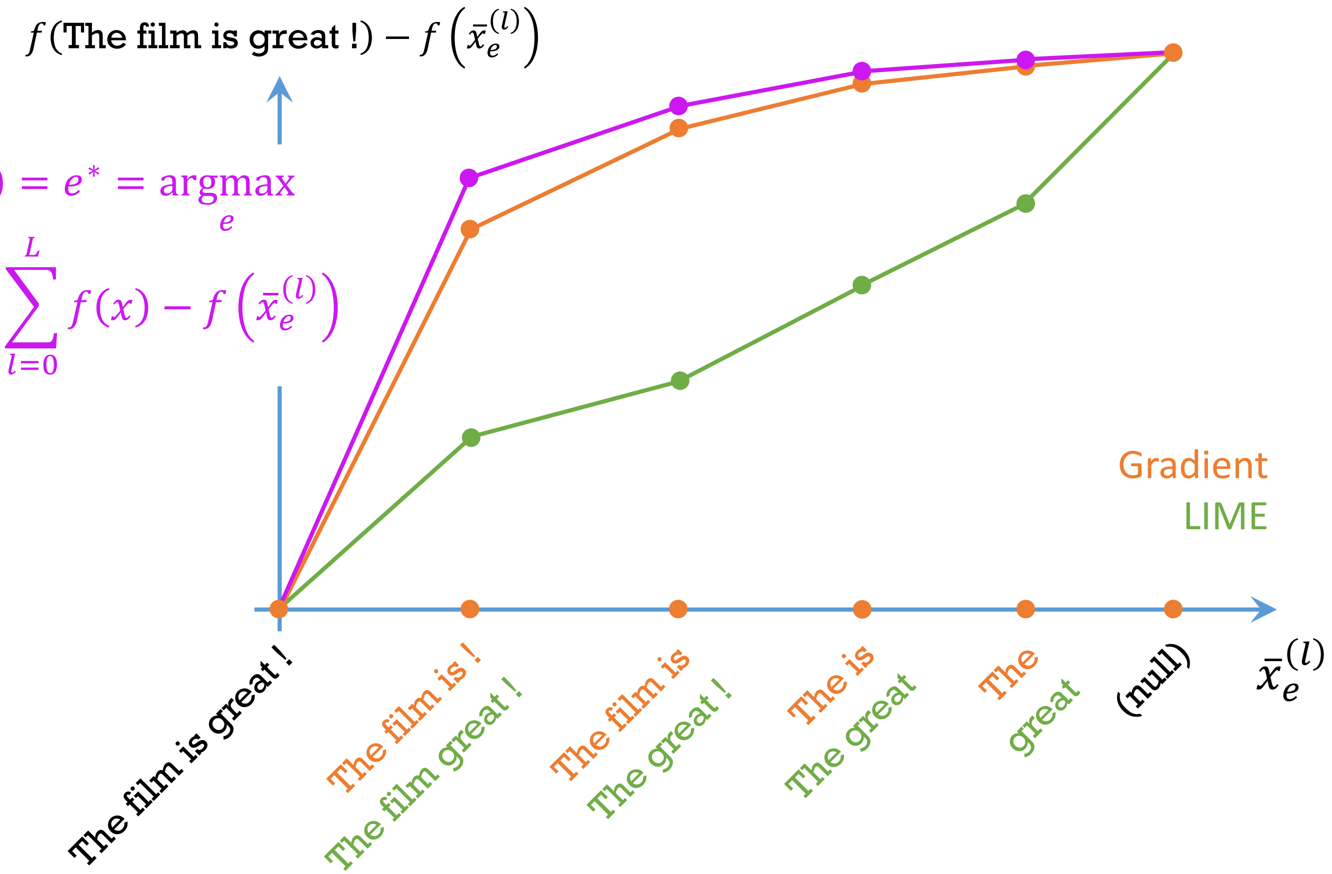
Explainer	Comp $\kappa \uparrow$	Suff $\sigma \downarrow$	Diff $\Delta \uparrow$
Grad	0.327	0.108	0.218
IntG	0.525	0.044	0.481
LIME	0.682	0.033	0.649
SHAP	0.612	0.034	0.578
Occl	0.509	0.040	0.469
E*	0.740	0.020	0.720
Random	0.218	0.212	0.006

Proxy Metrics for Explanation Quality

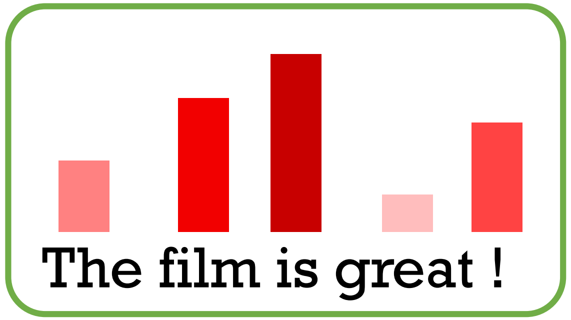
Slides and Resources

- Feature importance ⟺ model prediction change with feature removal

$$D_\kappa(x) = e^* = \underset{e}{\operatorname{argmax}} \frac{1}{L+1}\sum_{l=0}^{L} f(x) - f\left(\bar{x}_e^{(l)}\right)$$



Gradient



LIME

Evaluation on Other Aspects

Slides and Resources

- E* is competitive on other metrics

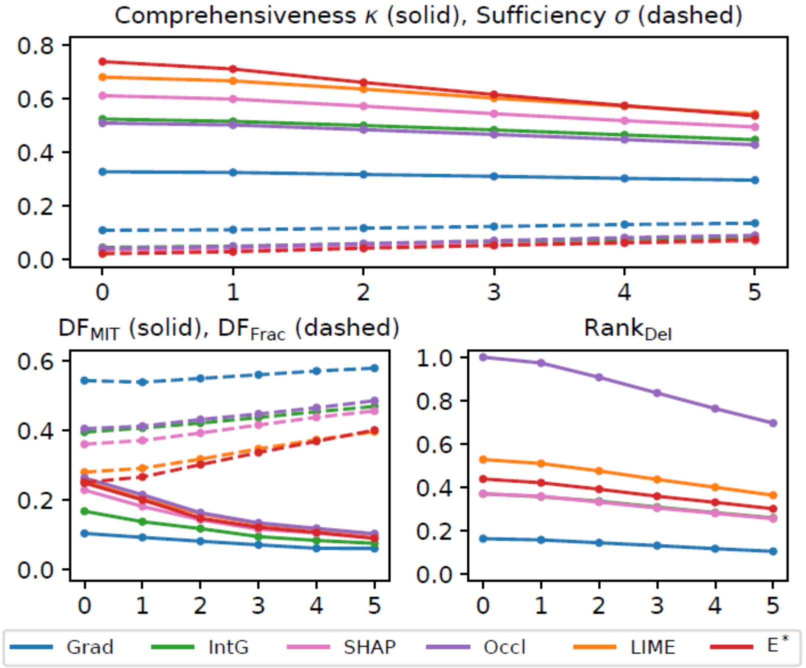
Explainer	$DF_{MIT}\uparrow$	$DF_{Frac}\downarrow$	$Rank_{Del}\uparrow$
Grad	10.5%	54.5%	0.162
IntG	16.9%	39.6%	0.369
LIME	25.5%	28.1%	0.527
SHAP	23.0%	36.1%	0.369
Occl	26.4%	40.6%	1.000
E*	25.0%	25.2%	0.438
Random	3.4%	72.3%	0.004

Evaluation on Other Aspects

Slides and Resources

- E* is competitive on other metrics
- E* is robust to perturbations

Explainer	$DF_{MIT}\uparrow$	$DF_{Frac}\downarrow$	$Rank_{Del}\uparrow$
Grad	10.5%	54.5%	0.162
IntG	16.9%	39.6%	0.369
LIME	25.5%	28.1%	0.527
SHAP	23.0%	36.1%	0.369
Occl	26.4%	40.6%	1.000
E*	25.0%	25.2%	0.438
Random	3.4%	72.3%	0.004



Time Efficiency

Slides and Resources

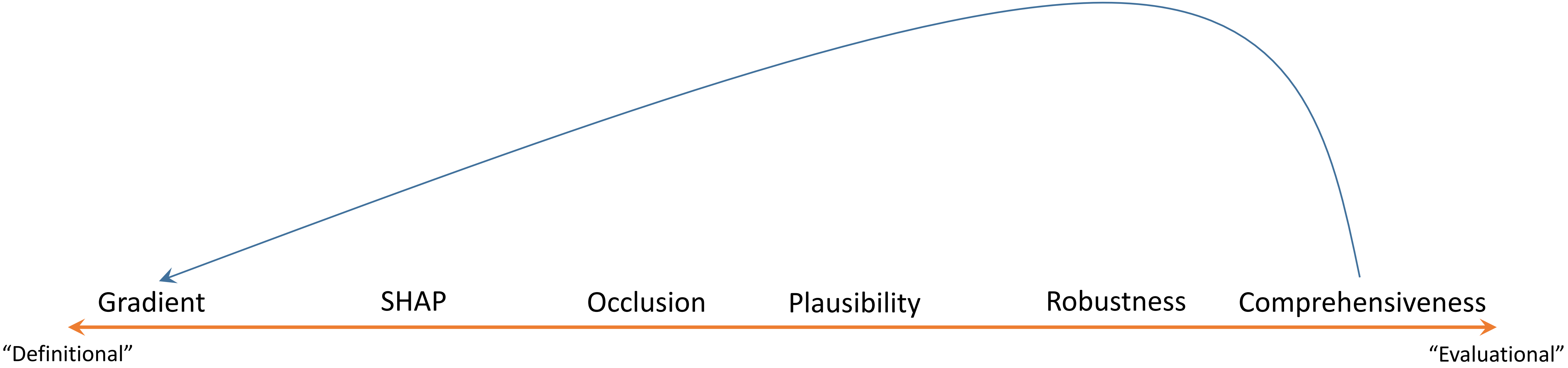
B	1	5	10	20	50	100	LIME
κ	0.717	0.731	0.734	0.736	0.739	0.740	0.682
σ	0.020	0.020	0.020	0.020	0.020	0.020	0.033
Δ	0.697	0.711	0.714	0.716	0.719	0.720	0.649
T	0.38	0.77	1.15	1.72	2.85	4.37	4.75

A New Paradigm of Developing Explainers?

Slides and Resources

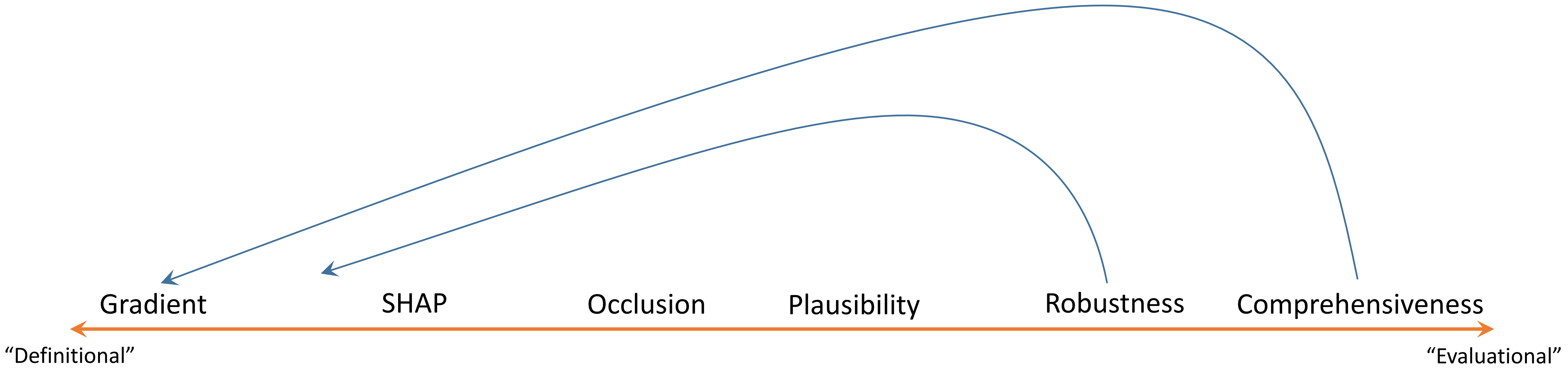
A New Paradigm of Developing Explainers?

Slides and Resources



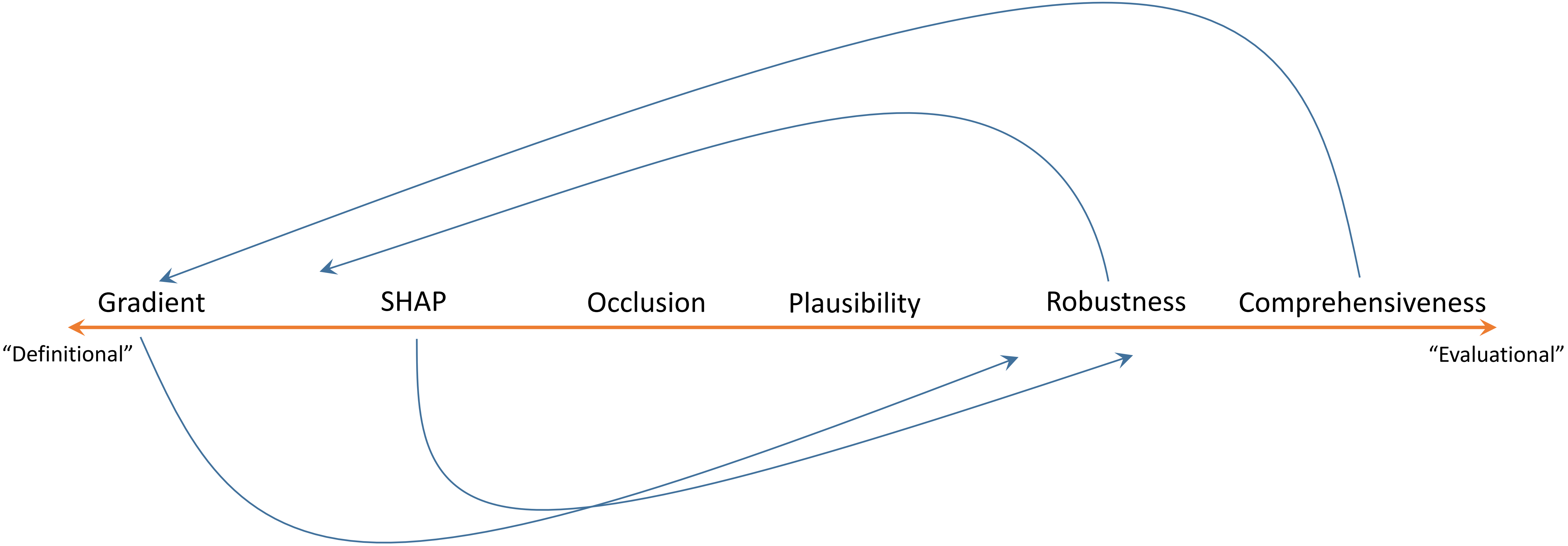
A New Paradigm of Developing Explainers?

Slides and Resources



A New Paradigm of Developing Explainers?

Slides and Resources



Solvability-Based Explainer

Slides and Resources

pip install solvex

https://yilunzhou.github.io/solvability/

Summary | NLP - Word | NLP - Sentence | CV - Grid Superpixel | CV - Custom Superpixel | Tabular

Read this tutorial as a Jupyter notebook here.

In this demo, we compute word-level explanations for the Huggingface `textattack/roberta-base-SST-2` model, also the setup presented in the paper.

We first load required packages and the RoBERTa model. Two classes are needed to compute the explanations. `BeamSearchExplainer` implements the beam search algorithm, and `*Masker` implements the feature masking. In this demo, we use `TextWordMasker` since we need to mask out individual words from a text input. The other demos showcase other `*Masker` s.

```
from solvex import BeamSearchExplainer, TextWordMasker
import torch
from transformers import AutoTokenizer, AutoModelForSequenceClassification
```

Solvability-Based Explainer

Slides and Resources

pip install solvex

Explained label: 1

Function value for label 1: 1.000

Contrary to other reviews, I have zero complaints about the service or the prices. I have been getting tire service here for the past 5 years now, and compared to my experience with places like Pep Boys, these guys are experienced and know what they're doing. Also, this is one place that I do not feel like I am being taken advantage of, just because of my gender. Other auto mechanics have been notorious for capitalizing on my ignorance of cars, and have sucked my bank account dry. But here, my service and road coverage has all been well explained - and let up to me to decide. And they just renovated the waiting room. It looks a lot better than it did in previous years.

Solvability-Based Explainer

Slides and Resources

pip install solvex

Explained label: 1

Function value for label 1: 1.000

Contrary to other reviews, I have zero complaints about the service or the prices. I have been getting tire service here for the past 5 years now, and compared to my experience with places like Pep Boys, these guys are experienced and know what they're doing. Also, this is one place that I do not feel like I am being taken advantage of, just because of my gender. Other auto mechanics have been notorious for capitalizing on my ignorance of cars, and have sucked my bank account dry. But here, my service and road coverage has all been well explained - and let up to me to decide. And they just renovated the waiting room. It looks a lot better than it did in previous years.

Explained label: 232. Function value: 0.159

Solvability-Based Explainer

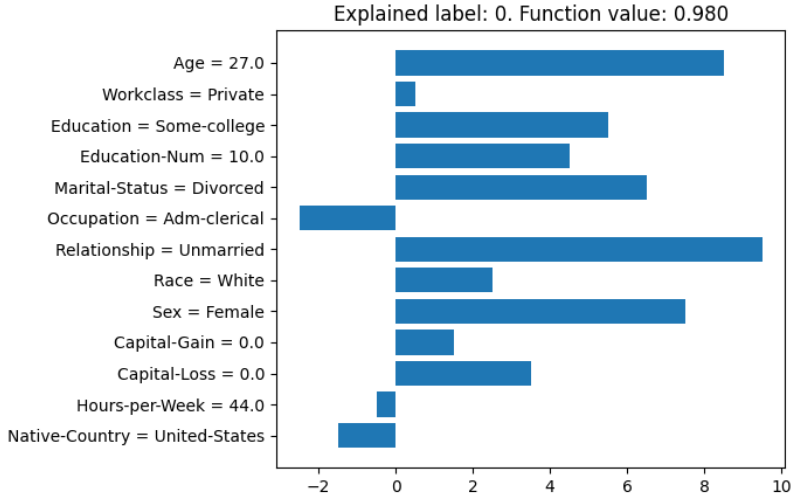
Slides and Resources

pip install solvex

Explained label: 1

Function value for label 1: 1.000

Contrary to other reviews, I have zero complaints about the service or the prices. I have been getting tire service here for the past 5 years now, and compared to my experience with places like Pep Boys, these guys are experienced and know what they're doing. Also, this is one place that I do not feel like I am being taken advantage of, just because of my gender. Other auto mechanics have been notorious for capitalizing on my ignorance of cars, and have sucked my bank account dry. But here, my service and road coverage has all been well explained - and let up to me to decide. And they just renovated the waiting room. It looks a lot better than it did in previous years.



Slides and Resources

References

- Definitions
  - Simonyan et al. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv 2014
  - Zeiler and Fergus. Visualizing and Understanding Convolutional Networks. ECCV 2014
  - Ribeiro et al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD 2016
  - Smilkov et al. SmoothGrad: Removing Noise by Adding Noise. arXiv 2017
  - Sundararajan et al. Axiomatic Attribution for Deep Networks. ICML 2017
  - Lundberg et al. A Unified Approach to Interpreting Model Predictions. NIPS 2017
- Evaluations
  - Samek et al. Evaluating the Visualization of What a Deep Neural Network Has Learned. T-NNLS 2016
  - Petsiuk et al. RISE: Randomized Input Sampling for Explanation of Black-box Models. BMVC 2018
  - Ghorbani et al. Interpretation of Neural Networks is Fragile. AAAI 2019
  - DeYoung et al. ERASER: A Benchmark to Evaluate Rationalized NLP Models. ACL 2020
  - Ross et al. Explaining NLP Models via Minimal Contrastive Editing (MiCE). ACL 2021 (Findings)
  - Zhou et al. Do Feature Attribution Methods Correctly Attribute Features? AAAI 2022
  - Zhou et al. ExSum: From Local Explanations to Model Understanding. NAACL 2022
- Duality
  - Zhou and Shah. The Solvability of Interpretability Evaluation Metrics. EACL 2023 (Findings)