Post Hoc, Local and Model-Agnostic Explanations

AAAI 2023 Tutorial on Trustworthy and Responsible AI

Yilun Zhou
MIT CSAIL & Amazon
Outline

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

<table>
<thead>
<tr>
<th>Task for DNN</th>
<th>Caption image</th>
<th>Recognise pneumonia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
<td>Describes green hillside as grazing sheep</td>
<td>Fails on scans from new hospitals</td>
</tr>
<tr>
<td>Shortcut</td>
<td>Uses background to recognise primary object</td>
<td>Looks at hospital token, not lung</td>
</tr>
</tbody>
</table>

(Geirhos et al. 2020)
Why Model Explanations?

Why is my model failing?
Because ... (debugging and diagnosis)

<table>
<thead>
<tr>
<th>Task for DNN</th>
<th>Caption image</th>
<th>Recognise pneumonia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
<td>Describes green hillside as grazing sheep</td>
<td>Fails on scans from new hospitals</td>
</tr>
<tr>
<td>Shortcut</td>
<td>Uses background to recognise primary object</td>
<td>Looks at hospital token, not lung</td>
</tr>
</tbody>
</table>

(Geirhos et al. 2020)
Why Model Explanations?

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)
Why Model Explanations?

How does my model predict this structure? Because ... (scientific discovery)

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)
The Two Major Axes of Interpretability

Global
Explains the model “more generally”

Local
Requires an input instance
The Two Major Axes of Interpretability

- **Global**
  - Explains the model "more generally"

- **Local**
  - Requires an input instance

- **Intrinsic**
  - Generated during model prediction

- **Post Hoc**
  - Generated by an external explainer after model prediction
The Two Major Axes of Interpretability

- **Global**: Explains the model “more generally”
- **Local**: Requires an input instance
- **Post Hoc**: Generated by an external explainer after model prediction
- **Intrinsic**: Generated during model prediction
The Two Major Axes of Interpretability

- **Global**
  - Explains the model “more generally”

- **Local**
  - Requires an input instance

- **Post Hoc**
  - Generated by an external explainer after model prediction

- **Intrinsic**
  - Generated during model prediction

**Neural additive model** (Agarwal et al., 2021)

Next presentation
The Two Major Axes of Interpretability

- **Global**
  - Explains the model “more generally”
  - 

- **Local**
  - Requires an input instance
  - 

- **Post Hoc**
  - Generated by an external explainer after model prediction
  - CNN filter visualization (Olah et al., 2017)
  - 

- **Intrinsic**
  - Generated during model prediction
  - Neural additive model (Agarwal et al., 2021)
  - 

Next presentation
The Two Major Axes of Interpretability

**Global**
Explains the model “more generally”

**Local**
Requires an input instance

**Intrinsic**
Generated during model prediction

**Post Hoc**
Generated by an external explainer after model prediction

- **CNN filter visualization** (Olah et al., 2017)
- **Neural additive model** (Agarwal et al., 2021)
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)](image-url)
Local *Post Hoc* Explanations

- **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

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

Counterfactual Explanation
Local *Post Hoc* Explanations

- 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)

Occlusion Saliency (Zeiler and Fergus, 2014)
Feature Attributions (Saliency Maps)

• Vanilla gradient: $s = \nabla_x f(x)$
Feature Attributions (Saliency Maps)

• Vanilla gradient: $s = \nabla_x f(x)$
• SmoothGrad: $s = \mathbb{E}_\epsilon [\nabla_x f(x + \epsilon)]; \quad \epsilon_i \text{ independent for every } x_i$
Feature Attributions (Saliency Maps)

- Vanilla gradient: \( s = \nabla_x f(x) \)
- SmoothGrad: \( s = \mathbb{E}_\epsilon [\nabla_x f(x + \epsilon)]; \ \epsilon_i \) independent for every \( x_i \)
Feature Attributions (Saliency Maps)

- 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: \( s = \int_0^1 \nabla_x f(x_0 + u(x - x_0)) \, du \)
Feature Attributions (Saliency Maps)

- 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: \( s = \int_0^1 \nabla_x f(x_0 + u(x - x_0)) \, du \)
- Occlusion: \( s_i = f(x) - f(x_{-i}) \)
Feature Attributions (Saliency Maps)

- **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**: \( s = \int_0^1 \nabla_x f(x_0 + u(x - x_0)) \, du \)
- **Occlusion**: \( s_i = f(x) - f(x_{-i}) \)
- **LIME**: \( f \sim s^T m + b \)

<table>
<thead>
<tr>
<th>Feature mask ( m )</th>
<th>Masked sentence</th>
<th>( f(\cdot) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 0, 0, 0</td>
<td>“”</td>
<td>0.49</td>
</tr>
<tr>
<td>0, 0, 0, 1</td>
<td>“beautiful”</td>
<td>0.9</td>
</tr>
<tr>
<td>0, 0, 1, 0</td>
<td>“and”</td>
<td>0.52</td>
</tr>
<tr>
<td>0, 0, 1, 1</td>
<td>“and beautiful”</td>
<td>0.91</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1, 1, 1, 0</td>
<td>“It’s good and”</td>
<td>0.89</td>
</tr>
<tr>
<td>1, 1, 1, 1</td>
<td>“It’s good and beautiful”</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Feature Attributions (Saliency Maps)

- 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: \( s = \int_0^1 \nabla_x f(x_0 + u(x - x_0)) \, du \)
- Occlusion: \( s_i = f(x) - f(x_{-i}) \)
- LIME: \( f \sim s^T m + b \)
- SHAP: \( s_i = \sum_{S \subseteq F \setminus \{i\}} \frac{(|S|! (|F| - |S| - 1)!)}{|F|!} \left[ f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \)
Feature Attributions (Saliency Maps)

- 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: \( s = \int_0^1 \nabla_x f(x_0 + u(x - x_0)) \, du \)
- 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)

\[ e = D(x) \in \mathbb{R}^L \]
Evaluating Explanations

Problem: don’t know how models work
Evaluating Explanations

Problem: don’t know how models work

Solution: develop explanation methods
Evaluating Explanations

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

Problem: don’t know how models work

Solution: develop explanation methods

Evaluation: compare with ground truth working mechanism
Proxy Metrics for Explanation Quality

• Feature importance $\iff$ model prediction change with feature removal
Proxy Metrics for Explanation Quality

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

The film is great!

Gradient
Proxy Metrics for Explanation Quality

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

The film is great!

Gradient
Proxy Metrics for Explanation Quality

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

The film is great!

Gradient

The film is great!
Proxy Metrics for Explanation Quality

• Feature importance $\iff$ model prediction change with feature removal

The film is great!
Proxy Metrics for Explanation Quality

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

*The film is great!*

Gradient

- The film is great!
- The film is!
- The film is
- The is
- The
- (null)
Proxy Metrics for Explanation Quality

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

The film is great!

Gradient
Proxy Metrics for Explanation Quality

• Feature importance $\iff$ model prediction change with feature removal

\[ f(\text{The film is great!}) - f(\tilde{x}_e(t)) \]

The film is great!
Proxy Metrics for Explanation Quality

- Feature importance $\iff$ model prediction change with feature removal
  
  $$f(\text{The film is great!}) - f(\tilde{x}_e(l))$$

---

The film is great! - The film is! - The film is - The - (null) - $\tilde{x}_e(l)$

Gradient

---

The film is great!
Proxy Metrics for Explanation Quality

- Feature importance \( \iff \) model prediction change with feature removal

\[
f(\text{The film is great!}) - f(\tilde{x}_e^{(l)})
\]

\[
\kappa(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\tilde{x}_e^{(l)})
\]

\(\tilde{x}_e^{(l)}\): 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

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

$$f (\text{The film is great!}) - f (\bar{x}^{(l)})$$

The film is great!

Gradient

The film is great!

LIME
Proxy Metrics for Explanation Quality

• Feature importance $\iff$ model prediction change with feature removal

$$f(\text{The film is great!}) - f(\tilde{x}_e^{(l)})$$
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

• 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.
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: \( q = E(x, e) \in \mathbb{R} \)
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

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?

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

\(X\)

\(Y\)

\(\rightarrow\) Crow

\(\rightarrow\) Crow

\(\rightarrow\) Sparrow

\(X\) : original input image

\(Y\) : original output label
Dataset Modification

\[ X \]

\[ Y \quad \rightarrow \quad \text{Crow} \]

\[ Y \quad \rightarrow \quad \text{Sparrow} \]

\[ \hat{Y} \sim \text{Ber}(0.5) \]

\[ X : \text{original input image} \]

\[ Y : \text{original output label} \]

\[ \hat{Y} : \text{modified output label} \]
Dataset Modification

\[ X \rightarrow \hat{X} \]

\[ Y \rightarrow \hat{Y} \sim \text{Ber}(0.5) \]

\( X \): original input image

\( Y \): original output label

\( \hat{Y} \): modified output label

\( \hat{X} \): modified input image
Dataset Modification

\[ X \rightarrow \hat{X} \]

\[ Y \rightarrow \hat{Y} \sim \text{Ber}(0.5) \]

Crow → Crow
Crow → Crow
Crow → Crow
Sparrow → Sparrow
Sparrow → Crow

Aggregate

Watermark on top
Watermark on bottom
Dataset Modification
Dataset Modification

\[
\% \text{ Attribution} = \frac{\sum A}{\sum A + \sum A}
\]
Evaluating Image Saliency Maps

% Attribution inside Injection Region

Feature Injection Region Size

- Random saliency map
- Optimal saliency map
Evaluating Image Saliency Maps

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

- SHAP Watermark
- SmoothGrad Watermark

% Attribution inside Injection Region

Feature Injection Region Size

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.

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

Mammogram

function $f$: pretrained on ROI patches

function $h$: whole image classification

Removal of heatmap

function $g$: using convolutional as top layers

{cancer, normal}

Whole image network feature map

Patch network feature map

Yilun Zhou  
Post Hoc, Local and Model-Agnostic Explanations  
MIT CSAIL & Amazon
“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)
Practical Implications

“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)

**Known spurious correlation**

**Some explainers don’t work**
Practical Implications

“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)

*Known spurious correlation*

*Some explainers don’t work*

*Unknown spurious correlation*

Can we trust these explainers?
Do the explanations correctly explain the model prediction logic?
Explanation Understandability

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

Computation trace: a fully correct explanation for the prediction
Correct but Not Understandable Explanations

Computation trace: a fully correct but totally not understandable explanation for the prediction
Input: As **shaky** as the plot is, Kaufman 's script is still **memorable** for some **great** one-liners.
Label: Positive
Model Prediction: Positive

SHAP score

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

General Rule
A positive sentiment word contributes very positively to the prediction.
Understanding a Sentiment Classifier

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

Label = Positive
Prediction = Positive

<table>
<thead>
<tr>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>memorable</td>
<td>0.48</td>
</tr>
<tr>
<td>great</td>
<td>0.37</td>
</tr>
</tbody>
</table>

A positive sentiment word contributes very positively to the prediction.
Understanding a Sentiment Classifier

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>memorable</td>
<td>0.48</td>
</tr>
<tr>
<td>great</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Label = Positive
Prediction = Positive

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>attractive</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Label = Negative
Prediction = Negative

A positive sentiment word contributes very positively to the prediction.
Understanding a Sentiment Classifier

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

- Label = Positive
- Prediction = Positive
- **memorable**: 0.48
- **great**: 0.37

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

- Label = Negative
- Prediction = Negative
- **attractive**: 0.00

A positive sentiment word contributes very positively to the prediction... **unless it is negated**
Understanding a Sentiment Classifier

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

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>memorable</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>great</td>
<td>0.37</td>
</tr>
</tbody>
</table>

A positive sentiment word contributes very positively to the prediction...

**unless it is negated**

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

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>attractive</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Daring, mesmerizing and exceedingly hard to forget.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label = Positive</td>
</tr>
<tr>
<td>Daring: 0.13</td>
</tr>
</tbody>
</table>

Supporting instance

Opposing instance
Understanding a Sentiment Classifier

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

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
<th>memorable</th>
<th>great</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>0.48</td>
<td>0.37</td>
</tr>
</tbody>
</table>

A positive sentiment word contributes very positively to the prediction... unless it is negated, or near another positive word.

**Daring, mesmerizing** and exceedingly hard to forget.

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
<th>Daring</th>
<th>mesmerizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>0.13</td>
<td>0.06</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
<th>attractive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Supporting instance

Opposing instance
Understanding a Sentiment Classifier

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

- **Label = Positive**
- **Prediction = Positive**

- **memorable:** 0.48
- **great:** 0.37

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

- **Label = Negative**
- **Prediction = Negative**

- **attractive:** 0.00

A positive sentiment word contributes very positively to the prediction...

**unless it is negated, or near another positive word**

**Daring, mesmerizing** and exceedingly hard to forget.

- **Label = Positive**
- **Prediction = Positive**

- **Daring:** 0.13
- **mesmerizing:** 0.06

Ranks among Williams’ **best** screen work.

- **Label = Positive**
- **Prediction = Positive**

- **best:** 0.05
Understanding a Sentiment Classifier

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

Label = Positive
Prediction = Positive

memorable: 0.48
great: 0.37

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

Label = Negative
Prediction = Negative

attractive: 0.00

A positive sentiment word **sometimes** contributes very positively to the prediction...

unless it is negated, or near another positive word

**Daring**, **mesmerizing** and exceedingly hard to forget.

Label = Positive
Prediction = Positive

Daring: 0.13
mesmerizing: 0.06

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

Label = Positive
Prediction = Positive

best: 0.05

Supporting instance
Opposing instance
Understanding a Sentiment Classifier

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

- Label = Positive
- Prediction = Positive

**memorable**: 0.48
**great**: 0.37

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

- Label = Negative
- Prediction = Negative

**attractive**: 0.00

A positive sentiment word **sometimes** contributes very positively to the prediction... unless it is negated, or near another positive word

**Daring, mesmerizing** and exceedingly hard to forget.

- Label = Positive
- Prediction = Positive

**Daring**: 0.13
**mesmerizing**: 0.06

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

- Label = Positive
- Prediction = Positive

**best**: 0.05

(many more counter-examples not shown)

Supporting instance

Opposing instance
Explanation Summary (ExSum)

Feature Attribution Explainer

Input 1 → Explanation 1
Input 2 → Explanation 2
Input 3 → Explanation 3

Over-generalization
Explanations Summary (ExSum)

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)

- Feature Attribution Explainer
  - Input 1
  - Input 2
  - Input 3
  - ... Input N

- Explanation 1
  - Explanation 2
  - Explanation 3
  - Explanation N

- ExSum

- Model Understanding 1
  - Model Understanding 2
  - ... Model Understanding M

- Metric Values:
  - Coverage
  - Validity
  - Sharpness

Systematic and quantitative model understanding
GUI for Developing ExSum Rules

ExSum Inspection

- Rule Union: (((R1 > R4) > R3) > (R5 > R6) > R7) 
  CF Without Rule 7: (((R1 > R4) > R3) > R5) > R6

**Rule Selection**
- Rule 1: negation
- Rule 2: highly positive adjectives have positive saliency
- Rule 3: highly negative adjectives have negative saliency
- Rule 4: highly positive words have positive saliency
- Rule 5: highly negative words have negative saliency
- Rule 6: Person names have small saliency
- Rule 7: Stop words have small saliency

**Metric Values**
- **Coverage**: Full 0.130 Full 0.605 Selected 0.475
- **Validity**: Full 0.911 Full 0.915 Selected 0.916
- **Sharpness**: Full 0.559 Full 0.269 Selected 0.195

**Example Visualization**
- **Sentence**: FEU
- **All**: Invalid
- **New Examples**

- **y=0**: If you collected all the moments of coherent dialogue, they still wouldn’t add up to the time required to boil a four-minute egg.
- **y=1**: Ranks among Williams’ best screen work.
- **y=1**: Like the film’s almost anthropologically detailed realization of early ‘80s suburbia, it’s significant without being overstated.
- **y=0**: It is a comedy that’s not very funny and an action movie that is not very thrilling (and an uneasy alliance, at that).
- **y=1**: A triumph, relentless and beautiful in its downbeat darkness.
- **y=0**: They felt like the same movie to me.
- **y=1**: The story feels more like a serious read, filled with heavy doses of always enticing Sayles dialogue.
- **y=1**: Behind the snow games and lovable Siberian huskies (plus one sheep dog), the picture hosts a parka-wrapped dose of heart.
- **y=1**: It is a film that will have people walking out halfway through, will encourage others to stand up and applaud, and will, undoubtedly, leave both camps engaged in a ferocious debate for years to come.
- **y=0**: De Niro may enjoy the same free ride from critics afforded to Clint Eastwood in the Bloodwork.
- **y=0**: The film might have been more satisfying if it had, in fact, been fleshed out a little more instead of going for easy smirks.
GUI for Developing ExSum Rules

$ 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

Correctness

Input → Local Explanation → Model Understanding

Understandability
The Many Faces of Understandability

- **Correctness**
- **Understandability**

**Input** → **Local Explanation** → **Model Understanding**

- **Human Alignment:**
  \[ e(x) \approx h(x) \text{ for human reasoning process } h(\cdot) \]

- **Robustness:**
  \[ ||e(x_1) - e(x_2)|| \leq \gamma ||x_1 - x_2|| \]

- **Similarity:**
  \[ ||e(x) - x|| \leq \epsilon \]
  (for counterfactual explanations)

- **Plausibility:**
  \[ p_D(e(x)) \geq \delta \]
The Many Faces of Understandability

Correctness

Understandability

Input → Local Explanation → Model Understanding

Different manifestations

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

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

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

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

(for counterfactual explanations)
Definition vs. Evaluation

**Definition**

Gradient
\[ D_g(x) = \nabla_x f(x) \]

**Evaluation**

Comprehensiveness
\[ E_k(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]
Definition vs. Evaluation

**Definition**

Gradient

\[ D_g(x) = \nabla_x f(x) \]

**Evaluation**

Comprehensiveness

\[ E_K(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]

Gradient

\[ E_g(x, e) = -\|\nabla_x f(x) - e\| \]
Definition vs. Evaluation

Definition

Gradient

\[ D_g(x) = \nabla_x f(x) \]

Comprehensiveness

\[ D_k(x) = \arg\max_e \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]

Evaluation

Comprehensiveness

\[ E_k(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]

Gradient

\[ E_g(x, e) = -||\nabla_x f(x) - e||\]
Definition-Evaluation Duality

**Definition**

- **Gradient**
  \[ D_g(x) = \nabla_x f(x) \]

- **Comprehensiveness**
  \[ D_{\kappa}(x) = \arg\max_e \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(x_e^{(l)}) \]

**Evaluation**

- **Comprehensiveness**
  \[ E_{\kappa}(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(x_e^{(l)}) \]

- **Gradient**
  \[ E_g(x, e) = -\|\nabla_x f(x) - e\| \]
Definition-Evaluation Duality

**Definitions**

- Gradient: $D_g(x) = \nabla_x f(x)$
- Comprehensiveness: $D_K(x) = \arg\max_e \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)})$

**Evaluation**

- Comprehensiveness: $E_K(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)})$
- Gradient: $E_g(x, e) = -||\nabla_x f(x) - e||$
Definition-Evaluation Duality

**Definition**

**Gradient**

\[ D_g(x) = \nabla_x f(x) \]

\[ D_g \gg \text{"ease" } D_\kappa \]

**Comprehensiveness**

\[ D_\kappa(x) = \arg\max_e \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]

**Evaluation**

**Comprehensiveness**

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

**Gradient**

\[ E_g(x,e) = -||\nabla x f(x) - e|| \]
Definition-Evaluation Duality

Definition

Gradient
\[ D_g(x) = \nabla_x f(x) \]

Comprehensiveness
\[ D_\kappa(x) = \text{argmax}_e \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]

Duality

Evaluation

Comprehensiveness
\[ E_\kappa(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\bar{x}_e^{(l)}) \]

Gradient
\[ E_g(x, e) = -||\nabla_x f(x) - e|| \]
Beam Search Results

\[
\kappa(x, e) = \frac{1}{L+1} \sum_{i=0}^{L} f(x) - f(\tilde{x}_e^{(i)})
\]

\[
e^* = \arg\max_e \kappa(x, e)
\]

**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. \(\text{beams} \leftarrow \{e^{(0)}\}\);
4. **for** \(l = 1, \ldots, L\) **do**
5. \(\text{beams} \leftarrow \bigcup_{e \in \text{beams}} \text{ext}(e, L - l + 1)\);
6. \(\text{beams} \leftarrow \text{choose_best}(\text{beams}, B)\);
7. **end**
8. \(e^* \leftarrow \text{choose_best}(\text{beams}, 1)\);
9. \(e^* \leftarrow \text{shift}(e^*)\);
10. **return** \(e^*\);
Beam Search Results

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

\[ e^* = \arg\max_e \kappa(x, e) \]

**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. \( \text{beams} \leftarrow \{e^{(0)}\} \);
4. for \( l = 1, ..., L \) do
5. \( \text{beams} \leftarrow \bigcup_{e \in \text{beams}} \text{ext}(e, L - l + 1) \);
6. \( \text{beams} \leftarrow \text{choose\_best(\text{beams}, B)} \);
7. end
8. \( e^* \leftarrow \text{choose\_best(\text{beams}, 1)} \);
9. \( e^* \leftarrow \text{shift}(e^*) \);
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

\[ \kappa(x, e) = \frac{1}{L + 1} \sum_{l=0}^{L} f(x) - f(\tilde{x}_e^{(l)}) \]

\[ e^* = \arg\max_e \kappa(x, e) \]

**Algorithm 1**: Beam search for finding \( e^* \).

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

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.

<table>
<thead>
<tr>
<th>Explainer</th>
<th>Comp ( \kappa ) ↑</th>
<th>Suff ( \sigma ) ↓</th>
<th>Diff ( \Delta ) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grad</td>
<td>0.327</td>
<td>0.108</td>
<td>0.218</td>
</tr>
<tr>
<td>IntG</td>
<td>0.525</td>
<td>0.044</td>
<td>0.481</td>
</tr>
<tr>
<td>LIME</td>
<td>0.682</td>
<td>0.033</td>
<td>0.649</td>
</tr>
<tr>
<td>SHAP</td>
<td>0.612</td>
<td>0.034</td>
<td>0.578</td>
</tr>
<tr>
<td>Occl</td>
<td>0.509</td>
<td>0.040</td>
<td>0.469</td>
</tr>
<tr>
<td><strong>E</strong></td>
<td><strong>0.740</strong></td>
<td><strong>0.020</strong></td>
<td><strong>0.720</strong></td>
</tr>
<tr>
<td>Random</td>
<td>0.218</td>
<td>0.212</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Proxy Metrics for Explanation Quality

• Feature importance $\iff$ model prediction change with feature removal

$$f(\text{The film is great!}) - f(\tilde{x}_e^{(l)})$$

$$D_K(x) = e^* = \arg\max_e$$

$$\frac{1}{L+1} \sum_{l=0}^{L} f(x) - f(\tilde{x}_e^{(l)})$$

The film is great!

The film is great!

Gradient

LIME

The film is great!

The film is great!
Evaluation on Other Aspects

- $E^*$ is competitive on other metrics

<table>
<thead>
<tr>
<th>Explainer</th>
<th>$DF_{MIT}$</th>
<th>$DF_{Frac}$</th>
<th>$Rank_{Del}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grad</td>
<td>10.5%</td>
<td>54.5%</td>
<td>0.162</td>
</tr>
<tr>
<td>IntG</td>
<td>16.9%</td>
<td>39.6%</td>
<td>0.369</td>
</tr>
<tr>
<td>LIME</td>
<td>25.5%</td>
<td>28.1%</td>
<td>0.527</td>
</tr>
<tr>
<td>SHAP</td>
<td>23.0%</td>
<td>36.1%</td>
<td>0.369</td>
</tr>
<tr>
<td>Occl</td>
<td>26.4%</td>
<td>40.6%</td>
<td>1.000</td>
</tr>
<tr>
<td>$E^*$</td>
<td>25.0%</td>
<td>25.2%</td>
<td>0.438</td>
</tr>
<tr>
<td>Random</td>
<td>3.4%</td>
<td>72.3%</td>
<td>0.004</td>
</tr>
</tbody>
</table>
Evaluation on Other Aspects

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

<table>
<thead>
<tr>
<th>Explainer</th>
<th>$DF_{MIT}$↑</th>
<th>$DF_{Frac}$↓</th>
<th>$Rank_{Del}$↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grad</td>
<td>10.5%</td>
<td>54.5%</td>
<td>0.162</td>
</tr>
<tr>
<td>IntG</td>
<td>16.9%</td>
<td>39.6%</td>
<td>0.369</td>
</tr>
<tr>
<td>LIME</td>
<td>25.5%</td>
<td>28.1%</td>
<td>0.527</td>
</tr>
<tr>
<td>SHAP</td>
<td>23.0%</td>
<td>36.1%</td>
<td>0.369</td>
</tr>
<tr>
<td>Occl</td>
<td>26.4%</td>
<td>40.6%</td>
<td>1.000</td>
</tr>
<tr>
<td>E*</td>
<td>25.0%</td>
<td>25.2%</td>
<td>0.438</td>
</tr>
<tr>
<td>Random</td>
<td>3.4%</td>
<td>72.3%</td>
<td>0.004</td>
</tr>
</tbody>
</table>
# Time Efficiency

<table>
<thead>
<tr>
<th>$B$</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>LIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$</td>
<td>0.717</td>
<td>0.731</td>
<td>0.734</td>
<td>0.736</td>
<td>0.739</td>
<td>0.740</td>
<td>0.682</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.020</td>
<td>0.020</td>
<td>0.020</td>
<td>0.020</td>
<td>0.020</td>
<td>0.020</td>
<td>0.033</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>0.697</td>
<td>0.711</td>
<td>0.714</td>
<td>0.716</td>
<td>0.719</td>
<td>0.720</td>
<td>0.649</td>
</tr>
<tr>
<td>$T$</td>
<td>0.38</td>
<td>0.77</td>
<td>1.15</td>
<td>1.72</td>
<td>2.85</td>
<td>4.37</td>
<td>4.75</td>
</tr>
</tbody>
</table>
A New Paradigm of Developing Explainers?

Gradient  SHAP  Occlusion  Plausibility  Robustness  Comprehensiveness

“Definitional”  “Evaluational”
A New Paradigm of Developing Explainers?

- Gradient
- SHAP
- Occlusion
- Plausibility
- Robustness
- Comprehensiveness

“Definitional” — “Evaluational”
A New Paradigm of Developing Explainers?

Gradient  
SHAP  
Oclusion  
Plausibility  
Robustness  
Comprehensiveness

“Definitional”  
“Evaluational”
A New Paradigm of Developing Explainers?

Gradient | SHAP | Occlusion | Plausibility | Robustness | Comprehensiveness

“Definitional” | “Evaluational”
Solvability-Based Explainer

pip install solvex

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

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 `*Maskers`.

```python
from solvex import BeamSearchExplainer, TextWordMasker
import torch
from transformers import AutoTokenizer, AutoModelForSequenceClassification
```
Solvability-Based Explainer

```
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

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
Contrary to other reviews, I have zero complaints about the service or the prices. I have been getting tired 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.
References

• Definitions
  • Zeiler and Fergus. Visualizing and Understanding Convolutional Networks. ECCV 2014
  • 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)