Evaluations and Methods for Explanation through Robustness Analysis

Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar,
Seungyeon Kim, Sanjiv Kumar, Cho-Jui Hsieh

Explanations for Machine Learning Models

Why Model Explanations?

- Machine learning models are deployed in many real-world applications, including high-stakes scenarios

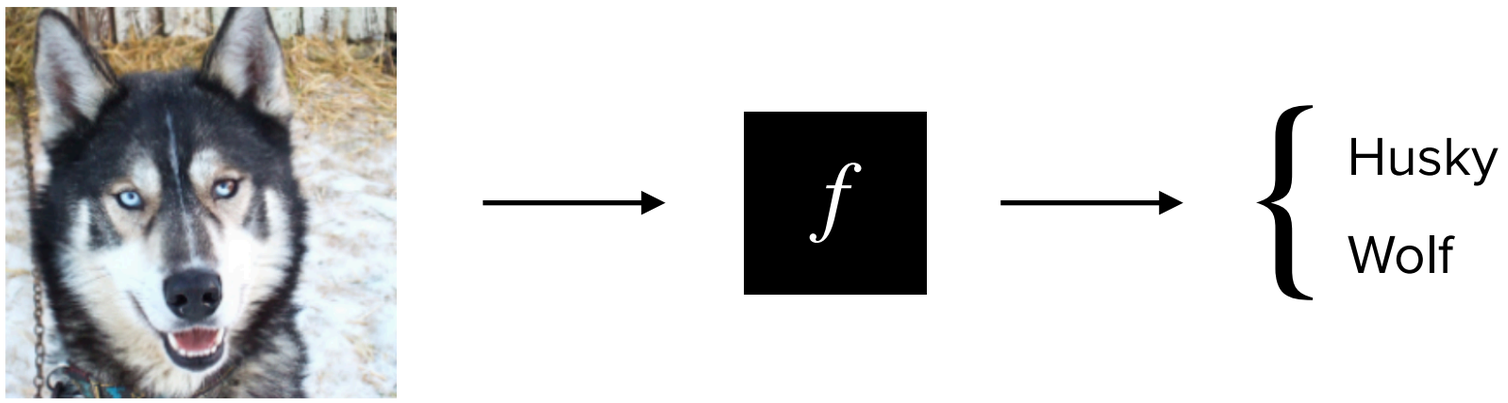
- Besides task performance such as accuracy, it is important to understand how the models work in order to establish user trust

Explanations for Machine Learning Models

Why Model Explanations?

- Machine learning models are deployed in many real-world applications, including high-stakes scenarios

- Besides task performance such as accuracy, it is important to understand how the models work in order to establish user trust



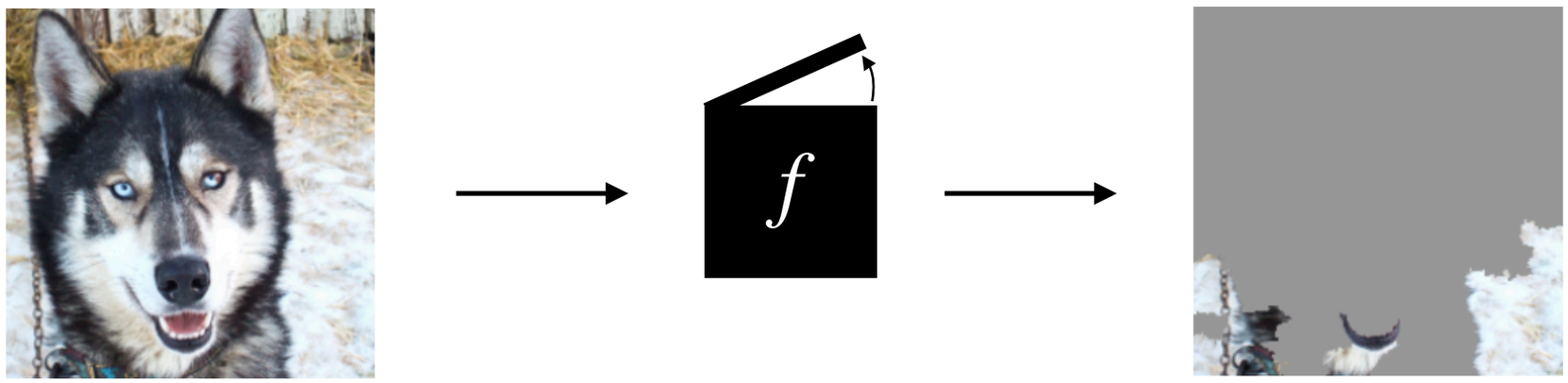
[Picture borrowed from Ribeiro et al., 2016]

Explanations for Machine Learning Models

Why Model Explanations?

- Machine learning models are deployed in many real-world applications, including high-stakes scenarios

- Besides task performance such as accuracy, it is important to understand how the models work in order to establish user trust



➡ **Explanations could help with understanding model trustability, fairness, weak points, etc.**

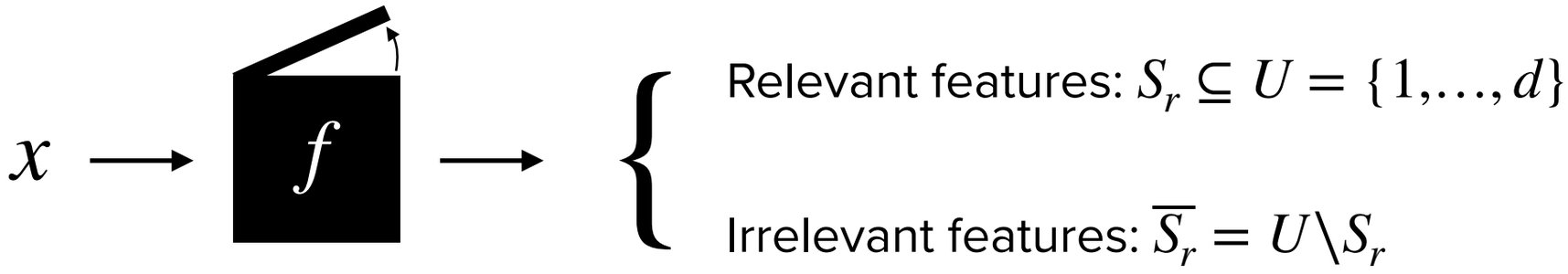
[Picture borrowed from Ribeiro et al., 2016]

Feature-based Explanations

Given an input example $x \in \mathbb{R}^d$, a model f, and its prediction $f(x)$:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \longrightarrow \boxed{f} \longrightarrow f(x)$$

Goal is to extract a compact set of relevant features with respect to the prediction:



Evaluating Feature Relevance

- **Necessity**: Removing relevant features from the input should lead to significant prediction change
  - If S_r is necessary for prediction: $f(x) - f(x_{U \setminus S_r})$ should be large
- **Sufficiency**: Removing irrelevant features, keeping only relevant features, should not lead to large prediction change
  - If S_r is sufficient for prediction: $f(x) - f(x_{S_r})$ should be small

Evaluating Feature Relevance

- **Necessity**: Removing relevant features from the input should lead to significant prediction change
  - If S_r is necessary for prediction: $f(x) - f(x_{U \setminus S_r})$ should be large
- **Sufficiency**: Removing irrelevant features, keeping only relevant features, should not lead to large prediction change
  - If S_r is sufficient for prediction: $f(x) - f(x_{S_r})$ should be small

➜ **Challenge:** Require ways to represent **feature removal**

- x_S is represented by $[x_s; x'_{\bar{s}}]$ where x' is some reference value
- Such reference value could introduce bias into the evaluation

Risk of Operationalizing Feature Removal

The use of reference value introduces **bias:**

- Features **close to the reference value** x' are likely to be considered **unimportant**:

$x_i = x_i'$ ↗ $f(x_{U\setminus i}) = f([x_{U\setminus i}; x_i']) = f(x)$ ⟶ Low Necessity

$x_i = x_i'$ ↘ $f(x_i) = f([x_i; x'_{U\setminus i}]) = f(x') \ll f(x)$ ⟶ Low Sufficiency

- Features **far away from the reference value** x' are more likely to be considered **important**:

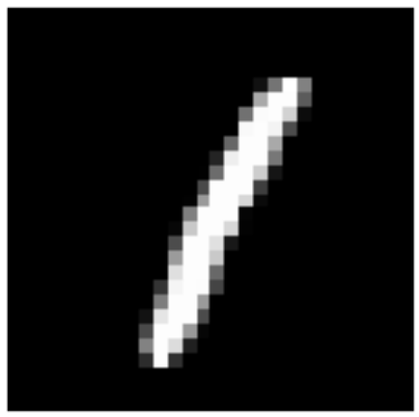
Large $|x_i - x_i'|$ ↗ Large $f(x) - f([x_{U\setminus i}; x_i'])$ ⟶ High Necessity

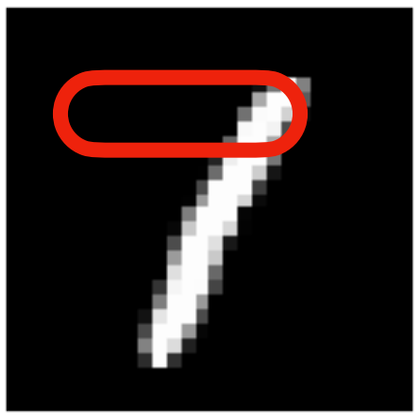
Large $|x_i - x_i'|$ ↘ Small $f(x) - f([x_i; x'_{U\setminus i}])$ ⟶ High Sufficiency

Risk of Operationalizing Feature Removal

Image classification with $x' = 0$:

- Black pixels will not be considered relevant
- However, they might be crucial in making the prediction
- E.g., turning the red circled area to white would possibly change the prediction from 1 to 7





Our Solution: Feature Robustness Analysis

Core Idea: from feature removal to **feature value perturbation**

- **Necessity:** ~~Removing~~ Perturbing the values of relevant features, fixing the irrelevant features, should lead to significant prediction change

- **Sufficiency:** ~~Removing~~ Perturbing the values of irrelevant features, fixing the relevant features, should not lead to large prediction change

- We propose to use **minimum adversarial perturbation norm** to quantify the influence of perturbations on the features

Robustness-based Twin Evaluation Criteria

Robustness on Feature Subset :

$$\epsilon_{x_S} = g(f, x, S) = \{\min_{\delta} \|\delta\|_p \;\; s.t. \;\; f(x+\delta) \neq f(x),\, \delta_{\overline{S}} = 0\}$$

- Given an explanation that partitions the input features into relevant feature set S_r and irrelevant feature set $\overline{S_r}$:
  - **Necessity** implies smaller robustness value on S_r

    Robustness-S_r $:= \epsilon_{x_{S_r}}$ (the smaller the better)

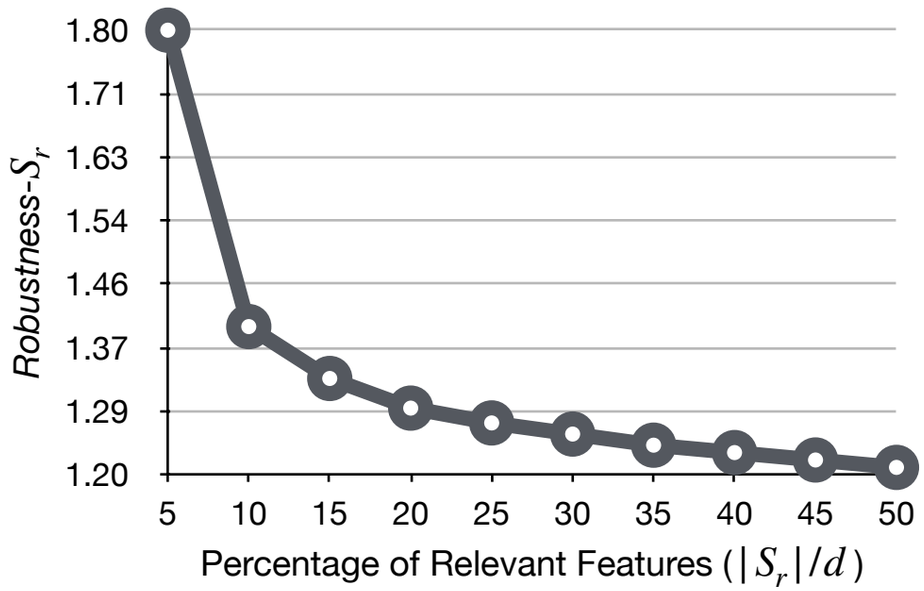
  - **Sufficiency** implies larger robustness value on $\overline{S_r}$

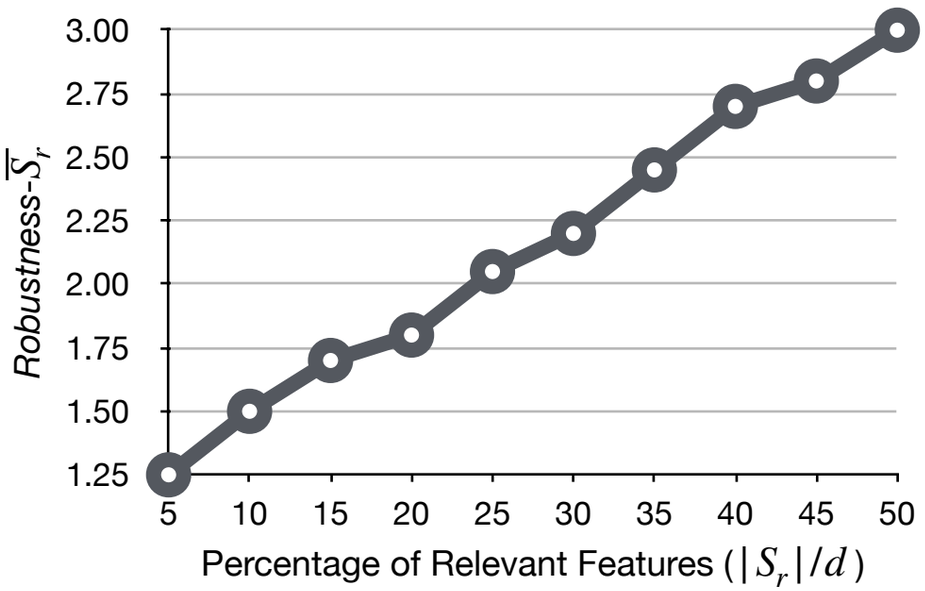
    Robustness-$\overline{S_r}$ $:= \epsilon_{x_{\overline{S_r}}}$ (the higher the better)
- Approximately compute ϵ_{x_S} by adversarial attacks

Robustness-based Twin Evaluation Criteria

Evaluation for Feature Importance Explanations

- Sort features by importance and provide top-K features as relevant set S_r
- Plot the evaluation curves of Robustness-S_r (-$\overline{S_r}$) by varying the size of S_r

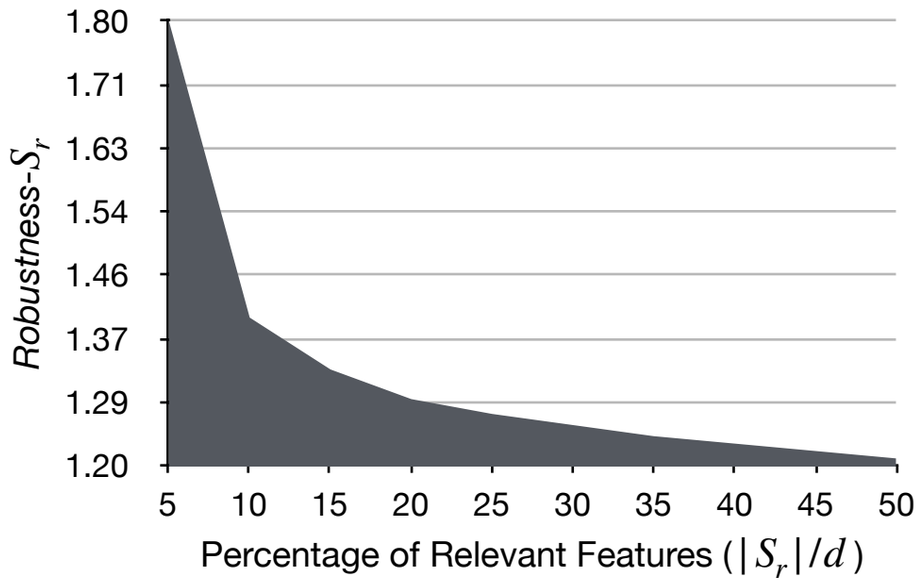


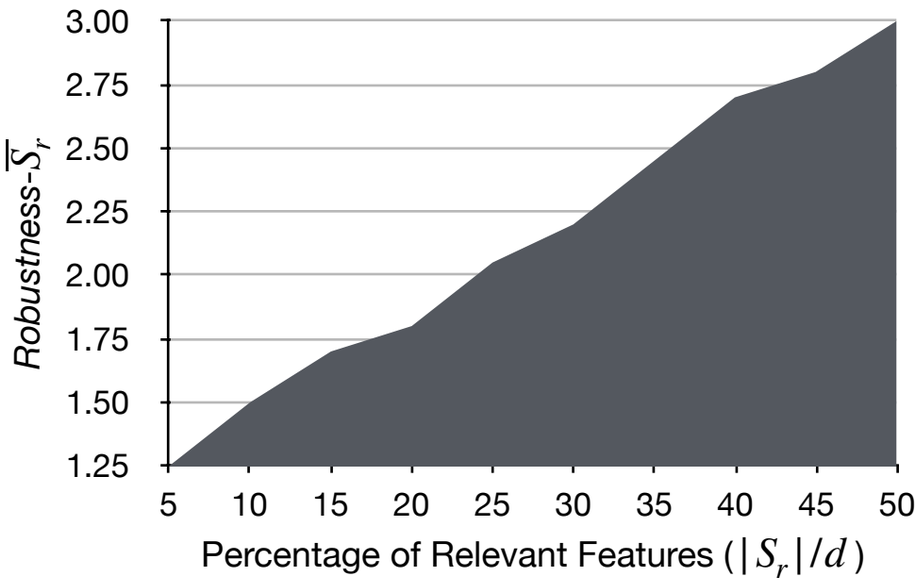


Robustness-based Twin Evaluation Criteria

Evaluation for Feature Importance Explanations

- Sort features by importance and provide top-K features as relevant set S_r
- Plot the evaluation curves of Robustness-S_r (-$\overline{S_r}$) by varying the size of S_r
- Smaller / Larger area under curve of Robustness-S_r / -$\overline{S_r}$ indicates better feature attribution ranking





Contrastive Explanation by Targeted Attack

Untargeted Adversarial Robustness:

$$\epsilon_{x_S} = g(f, x, S) = \{\min_{\delta} \|\delta\|_p \;\; s.t. \;\; f(x+\delta) \neq f(x),\, \delta_{\overline{S}} = 0\}$$

- Relevant features that lead to the current prediction $f(x)$

Targeted Adversarial Robustness:

$$\epsilon_{x_S,t} = g(f, x, S) = \{\min_{\delta} \|\delta\|_p \;\; s.t. \;\; f(x+\delta) = t,\, \delta_{\overline{S}} = 0\}$$

where t is the targeted class

- Relevant features that lead to its current prediction $f(x)$ but not class t
- Answers the question “Why an example is classified as A but not B?”

New Explanation Optimizing the Evaluation

- Searching for an optimal set of relevant features S_r , under a cardinality constraint, leads to the following set of optimization problems:

Minimize Robustness-S_r

$$\arg\min_{S_r \subseteq U} g(f, x, S_r) \quad s.t. \quad |S_r| \leq K$$

Maximize Robustness-$\overline{S_r}$

$$\arg\max_{S_r \subseteq U} g(f, x, \overline{S_r}) \quad s.t. \quad |S_r| \leq K$$

- Directly solving these problems is challenging given that computing $g(\cdot)$ is itself NP-hard, which is further exacerbated by the discrete input constraint

New Explanation Optimizing the Evaluation

Naive Greedy Algorithm (Greedy):

1. Initialize $S_r^0 = \emptyset$
2. $S_r^{t+1} = S_r^t \cup i$ where i is selected by:

   Robustness-S_r : $\arg\min_i g(f, x, S_r^t \cup i)$ / Robustness-$\overline{S_r}$: $\arg\max_i g(f, x, \overline{S_r^t \cup i})$

3. Repeat Step 2 until $|S_r| = K$

➜ **Downside:** Feature interactions are ignored

- Features seem irrelevant when evaluated independently might nonetheless be relevant when evaluated simultaneously

New Explanation Optimizing the Evaluation

Greedy by Set Aggregation Score (Greedy-AS):

- **Key Idea:** Iteratively choose features based on their expected contribution to the objective $g(\cdot)$ when added to S_r, along with a random subset of other unchosen features

- Measure the aggregated contribution score via a linear regression:

$$w^t = \arg\min_{w} \min_{c} \sum_{S \in \mathcal{P}(\overline{S_r^t})} ((w^T b(S) + c) - v(S_r^t \cup S))^2$$

  where $b : \mathcal{P}(\overline{S_r^t}) \rightarrow \{0,1\}^{|\overline{S_r^t}|}$ projects S into its binary vector form and $v(\cdot)$ is the objective function of interest

➡ w^t corresponds to the unchosen features' expected contribution to the objective when included into S_r

- At each step t, choose features that are expected to contribute the most

Quantitative Evaluation of Greedy-AS

Evaluation under Robustness-S_r / $\overline{S_r}$

Table 1: AUC of Robustness-$\overline{S_r}$ and Robustness-S_r for various explanations on different datasets. The higher the better for Robustness-$\overline{S_r}$; the lower the better for Robustness-S_r.

Datasets	Explanations	Grad	IG	EG	SHAP	LOO	BBMP	CFX	Random	Greedy-AS
MNIST	Robustness-$\overline{S_r}$	88.00	85.98	93.24	75.48	74.14	78.58	69.88	64.44	**98.01**
	Robustness-S_r	91.72	91.97	91.05	101.49	104.38	176.61	102.81	193.75	**82.81**
ImageNet	Robustness-$\overline{S_r}$	27.13	26.01	26.88	18.25	22.29	21.56	27.12	17.98	**31.62**
	Robustness-S_r	45.53	46.28	48.82	60.02	58.46	158.01	46.10	56.11	**43.97**
Yahoo!Answer	Robustness-$\overline{S_r}$	1.97	1.86	1.96	1.81	1.74	-	1.95	1.71	**2.13**
	Robustness-S_r	2.91	3.14	2.99	3.34	4.04	-	2.96	7.64	**2.41**

➡ Greedy-AS effectively optimizes the proposed criteria

Evaluation under Insertion/Deletion

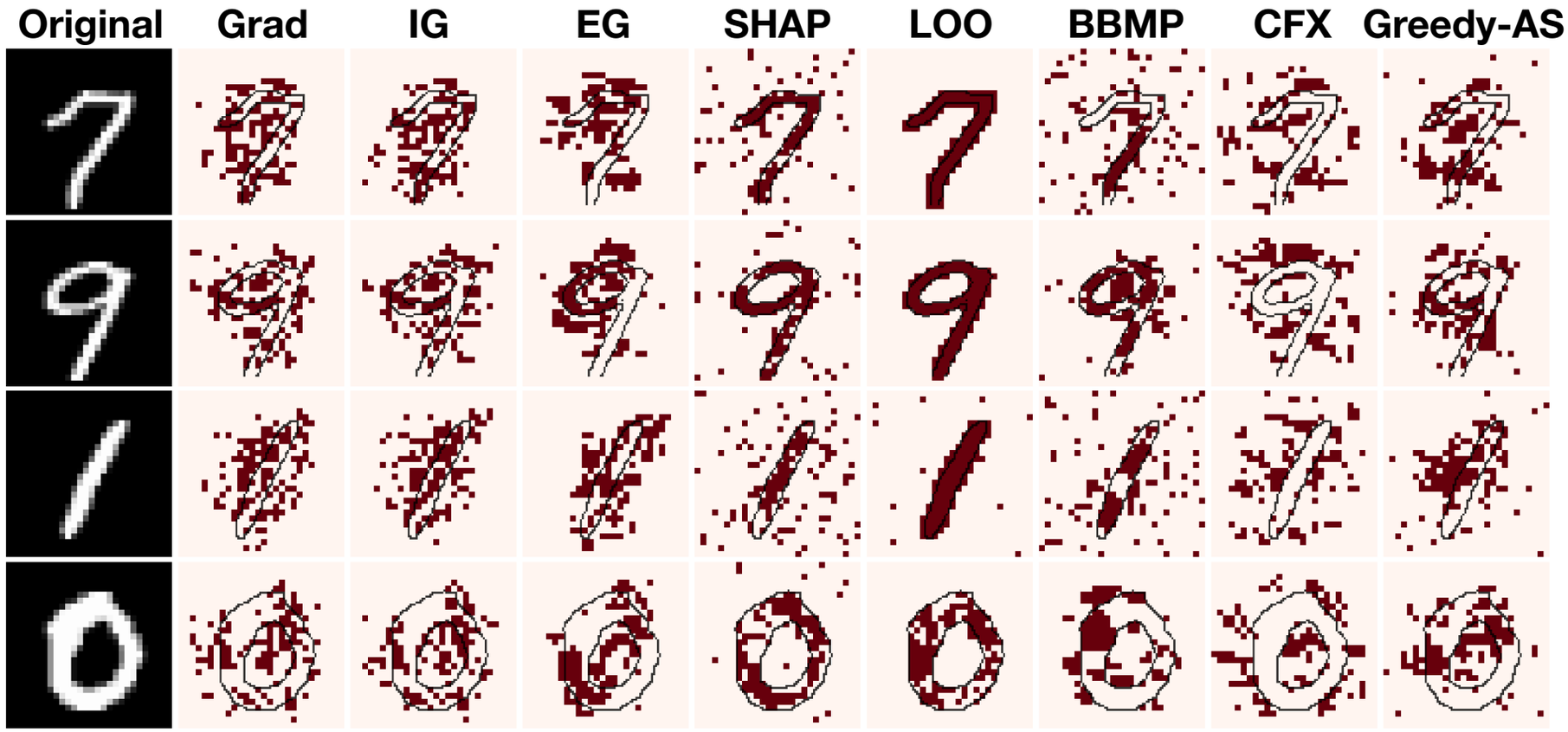
Table 2: AUC of the Insertion and Deletion criteria for various explanations on different datasets. The higher the better for Insertion; the lower the better for Deletion.

Datasets	Explanations	Grad	IG	EG	SHAP	LOO	BBMP	CFX	Random	Greedy-AS
MNIST	Insertion	174.18	177.12	228.64	125.93	121.99	108.97	102.05	51.71	**270.75**
	Deletion	153.58	150.90	113.21	213.32	274.77	587.08	137.69	312.07	**94.24**
ImageNet	Insertion	86.16	109.94	150.81	28.06	63.90	135.98	97.33	31.73	**183.66**
	Deletion	276.78	256.51	244.88	**143.27**	290.10	615.13	281.12	314.82	219.52
Yahoo!Answers	Insertion	0.06	0.06	0.20	0.07	0.18	-	0.05	0.10	**0.21**
	Deletion	2.57	2.96	2.07	2.23	2.07	-	2.35	2.63	**1.56**

➡ Greedy-AS also performs favorably on a set of existing popular criteria

Qualitative Evaluation of Greedy-AS

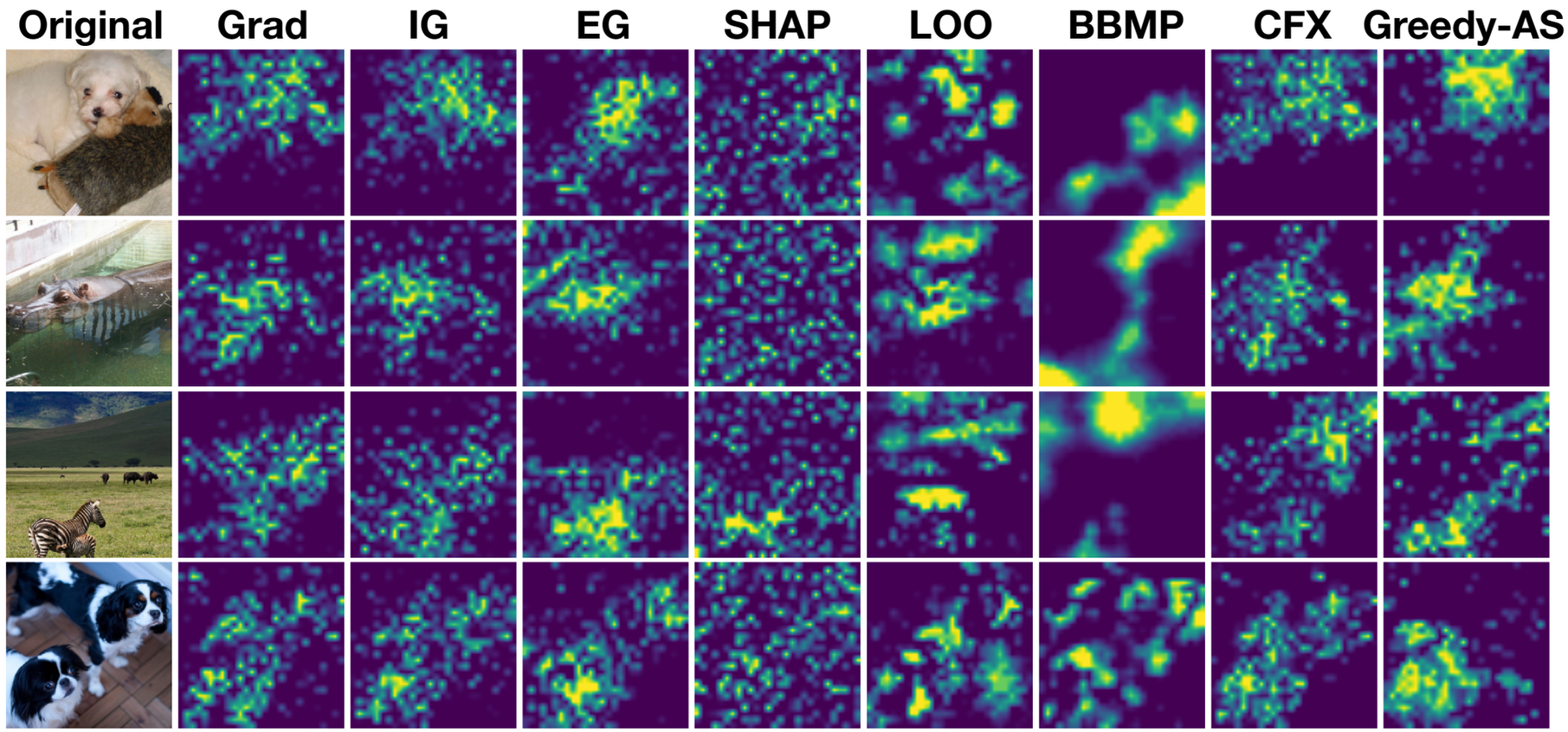
Explanations on MNIST



➜ Greedy-AS highlights both important white and black pixels, while existing explanations tend to focus more on the white pixels

Qualitative Evaluation of Greedy-AS

Explanations on ImageNet



➡ Greedy-AS focuses more compactly on the actual objects being classified

Qualitative Evaluation of Greedy-AS

Explanations on Yahoo!Answers

Input	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
Grad	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
IG	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
EG	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
SHAP	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
LOO	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
CFX	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
Greedy-AS	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?
Anchor	Ronaldinho and kaka are my favorite players out there. why did they replace them? I completely missed that part. Do they say why the switched them?

Most Relevant **Less Relevant**

➡ Top-5 keywords selected by Greedy-AS are all related to the label “Sports”

Qualitative Evaluation of Greedy-AS

Targeted Explanations on MNIST

➜ Highlighted features change meaningfully when the targeted class changes

Conclusion

- We define new evaluation criteria for feature based explanations by leveraging robustness analysis
  - This reduces the bias inherent in other recent evaluation measures that focus on “removing features”
- We design efficient algorithms to generate explanations that optimize the proposed criteria
  - We demonstrate the effectiveness and interpretability of our proposed explanation on image and language datasets

Thank You!

Evaluations and Methods for Explanation through Robustness Analysis

Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar,

Seungyeon Kim, Sanjiv Kumar, Cho-Jui Hsieh