Do Explains Reflect Decisions?

A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms

Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev,
Michael St. Jules, Xiao Yu Wang, Alexander Wong
https://arxiv.org/abs/1910.07387

UNIVERSITY OF WATERLOO

VIPlab
vip.uwaterloo.ca

DARWIN AI

EXPLAINABILITY EXAMPLE:

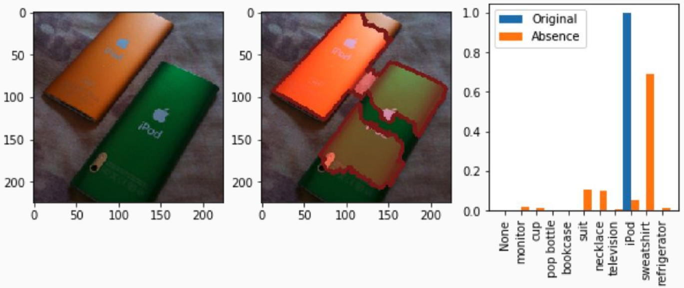
- An edge-case in our autonomous vehicle work
- Predict to turn left, if sky is purple

The Solution?

- Problem: How can we know a network is using the right features?

Explainability Algorithms!

What Does Explainability Mean?



Critical factors highlighted with red contour

For this talk:

- **Critical Factors:** the most important subset features of a given input for the model's prediction
- **Saliency Map:** the heat map where values for each feature representing how important that feature is
- Thinking in counterfactuals: what would the output be, if it weren't for these features.

Should We Blindly Trust Explainability Algorithms?

- **No**, at least not at the moment.
- Problem with explainability algorithms
  - Trade-off between stability and efficiency
  - Inconsistent explanation
  - Reliance on human verification
    - Time-consuming
    - Subjective

Evaluation of Explainability Algorithms

Solution to reliance on human verifications?

- **Machine-centric Evaluation Metric**
  - Define quality of explanations **Quantitatively**
  - Compare quality of explanations between algorithms

Terminologies

- Denote the following:
  - *N*: Neural network
  - *x*: Input to the network
  - *y*: Prediction of the network
  - *z*: Confidence of the prediction
  - *c*: Critical factors

$$(y, z) = N(x)$$

$$c = M(x, N) \subseteq x$$

$$x' = x - c$$

x without *c*, by setting the critical factors to 0

$$(y', z') = N(x')$$

Our Metric: Impact Score

Two types of impact are considered:

- Decision-level impact:
  - *y' ≠ y* => Decision changes without the critical factors
- Confidence-level impact:
  - *z' ≤ z - τ* => Confidence drops by a given constant (τ = 0.5 in our experiment)

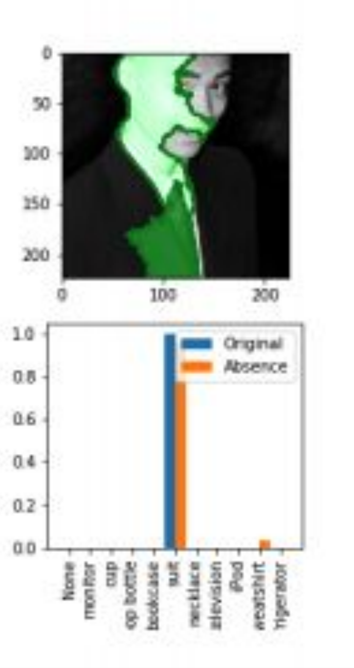
- Strict Impact Score I_{strict}:

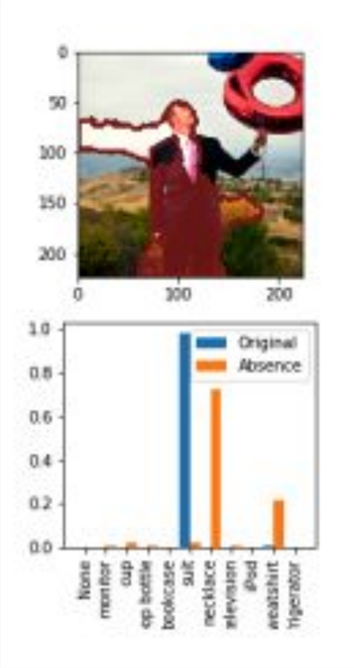
$$I_{strict} = \frac{1}{n}\sum_{i=1}^{n}(y_i' \neq y_i)$$

- Impact Score *I*:

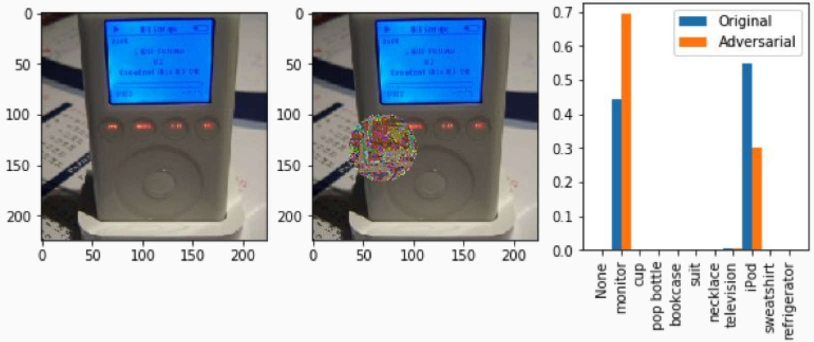
$$I = \frac{1}{n}\sum_{i=1}^{n}((y_i' \neq y_i) \vee (z_i' \leq \tau z_i))$$

Our Metric: Impact Score





Our Metric: Impact Coverage



$$c = M(x, N) \subseteq x$$

where x has an adversarial patch as in (Brown et al., 2017) applied to it.

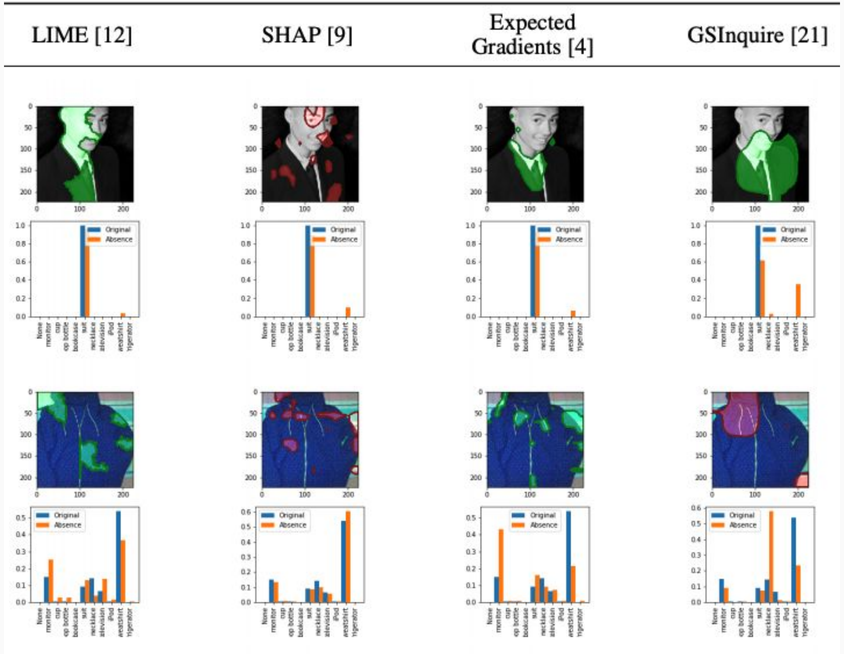
α: the adversarial patch in *x*

- Apply adversarial patch as in (Brown et al., 2017) on the input image
- **Impact Coverage:**
  - The mean IOU between adversarial patches and the critical factors

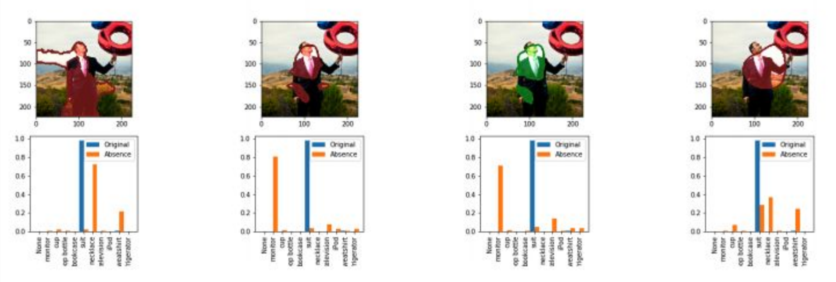
$$I_{coverage} = mean\ IOU = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i \cap c_i|}{|a_i \cup c_i|}$$

Our Metric: Impact Coverage

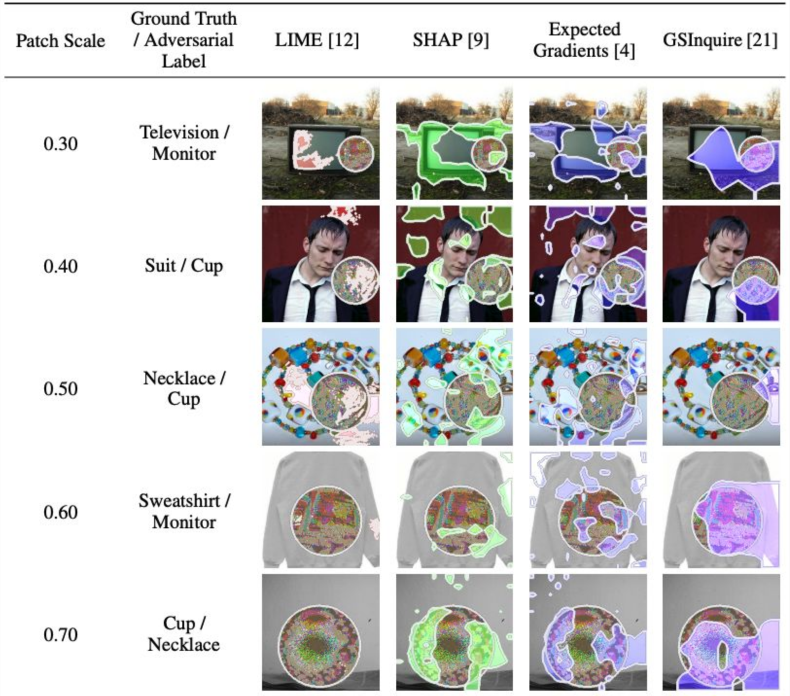
Experiment 1:



Method	I	I_{Strict}
LIME [12]	38.05%	35.12%
SHAP [9]	44.15%	40.24%
Expected Gradients [4]	51.22%	47.80%
GSInquire [21]	76.10%	50.73%



Experiment 2:



Scale	LIME [12]			SHAP [9]			Expected Gradient [4]			GSInquire [21]		
	$I_{coverage}$	I	I_{strict}	$I_{coverage}$	I	I_{strict}	$I_{coverage}$	I	I_{strict}	$I_{coverage}$	I	I_{strict}
0.3	0.64%	9.70%	9.80%	3.53%	40.41%	41.32%	2.57%	36.00%	36.80%	13.90%	66.90%	68.00%
0.4	1.53%	9.90%	10.00%	3.33%	36.73%	37.54%	2.31%	35.00%	35.40%	19.24%	64.50%	65.80%
0.5	0.67%	8.70%	8.80%	3.08%	36.28%	36.62%	2.09%	39.20%	39.40%	20.02%	66.90%	67.80%
0.6	0.37%	10.50%	10.60%	3.04%	38.20%	38.78%	1.88%	39.00%	39.40%	19.09%	67.20%	67.90%
0.7	0.41%	10.80%	10.80%	2.87%	43.16%	43.61%	1.80%	42.80%	43.20%	17.29%	68.90%	69.70%

Conclusion and Future Work

- Some of the most popular and widely-used explainability methods may produce explanations that may not be as reflective for the decision
- Extend this framework to different task domains such as natural language processing and audio understanding.