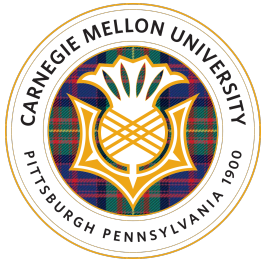
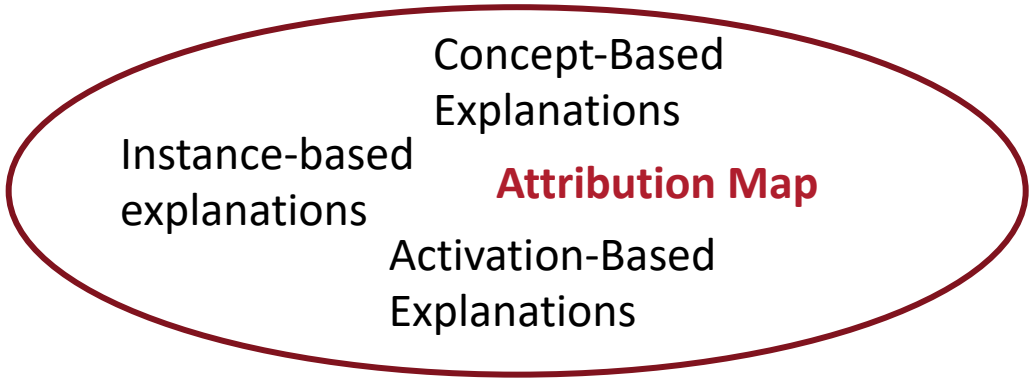
IEEE CVPR2020 WORKSHOP ON
FAIR, DATA EFFICIENT AND TRUSTED COMPUTER VISION

Interpreting Interpretations: Organizing Attribution Methods by Criteria

Zifan Wang, Piotr Mardziel, Anupam Datta, Matt Fredrikson
Accountable System Lab (https://fairlyaccountable.org/)
Carnegie Mellon University

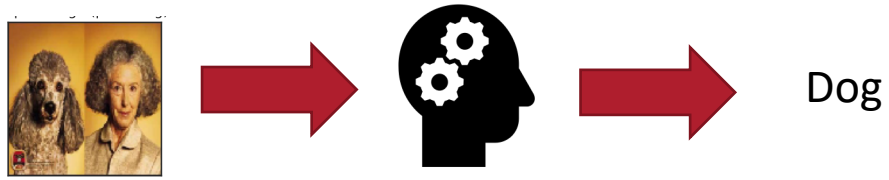


Introduction



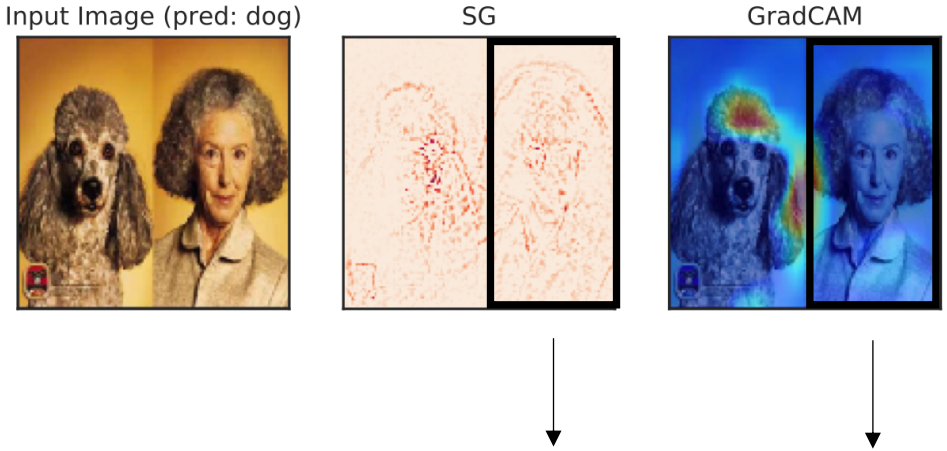
Input $x = [x_0, x_1, \dots, x_{n-1}]$

Attribution Map $z = [z_0, z_1, \dots, z_{n-1}]$



Attribution Map : Find the most important features in the input towards the prediction

Attribution Maps may not agree with each other



The lady **IS** or **IS NOT** used by the model to predict “dog” class

Background

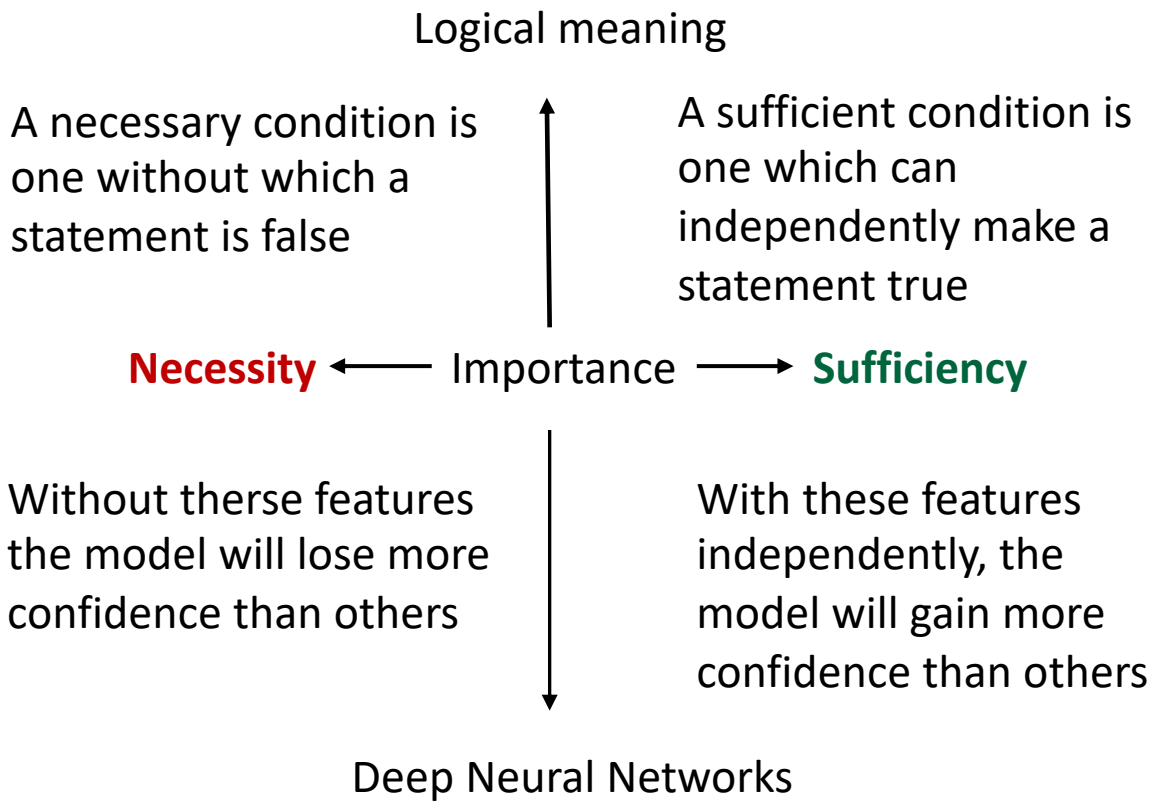
- Goal of this paper：

Evaluate different attribution methods with **numerical analysis** to answer which attribution method is better at what extent ?

- Recall:

Attribution Map : Find the most **important** features in the input towards the prediction

- Decompose the **importance**



Methods

Quantifying Necessity & Sufficiency

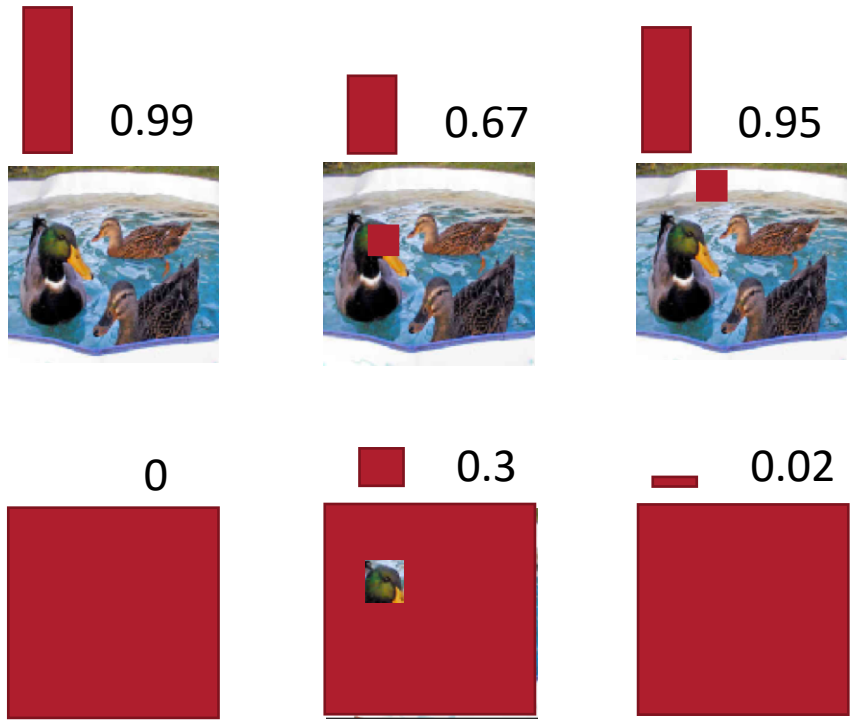
- Criteria One: Ordering

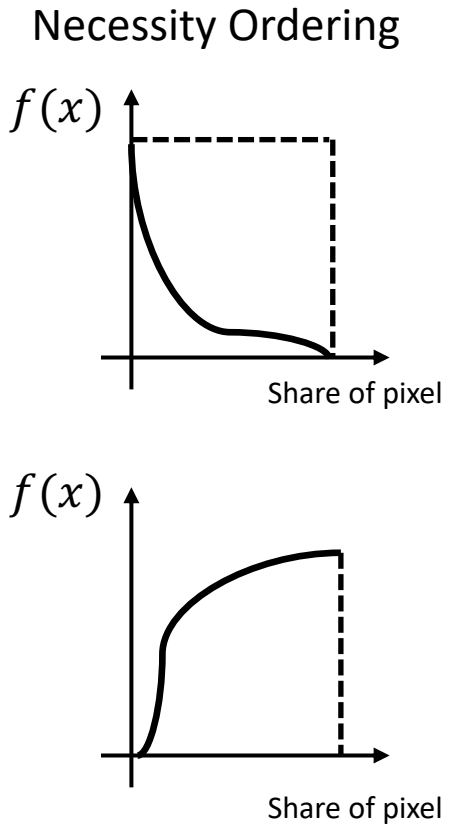
Scores → Rank Order

Quantify the ordering: **Modify** features from top rank orders to the bottom.

Modify → Ablate (Necessity)

Modify → Add (Sufficiency)





Similar Metrics

Necessity Ordering： AOPC [Samek et al. 2017],
AUC [Binder et al. 2016],
MoRF [Ancona et al. 2017]

Sufficiency Ordering：LeRF [Ancona et al. 2017]
Average % Dorp [Chattopadhyay et al. 2017]

Methods

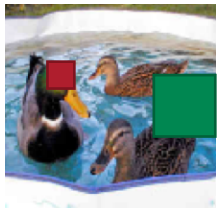
Quantifying Necessity & Sufficiency

- Criteria One: Ordering

However, there is more than rank orders that attribution scores offer, and the actual values are overlooked.

Methods

Quantifying Necessity & Sufficiency



Sum (■) = Sum (■)

- Criteria Two: Proportionality

Motivations:

- Magnitude also captures information
- Interpret attribution scores linearly: given x_1, x_2, if $z_1 = 2z_2$, then x_1 is expected to be twice important than x_2 is.

Equally important

Sensitivity-N [Ancona et al. 2017]

Randomly choose regions -->
Loosing rank order information

Property under a given ordering:	Method	
Ours: Choose ■ from Top 1 … N	TPN	**Necessity**
Ours: Choose ■ from Top N … 1	TPS	**Sufficiency**

Methods

Quantifying Necessity & Sufficiency

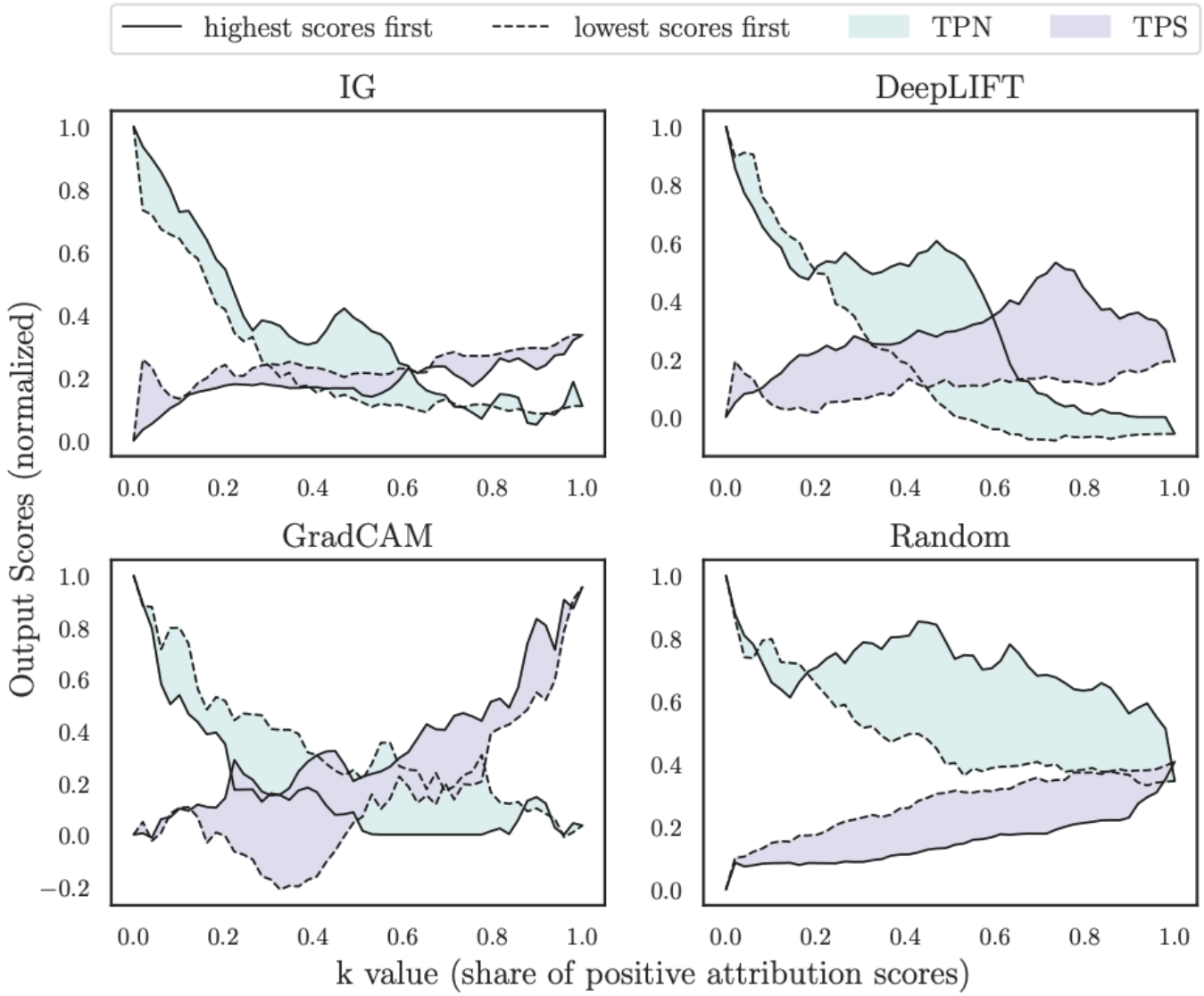
- Criteria Two: Proportionality

TPN measures Necessity

TPS measures Sufficiency

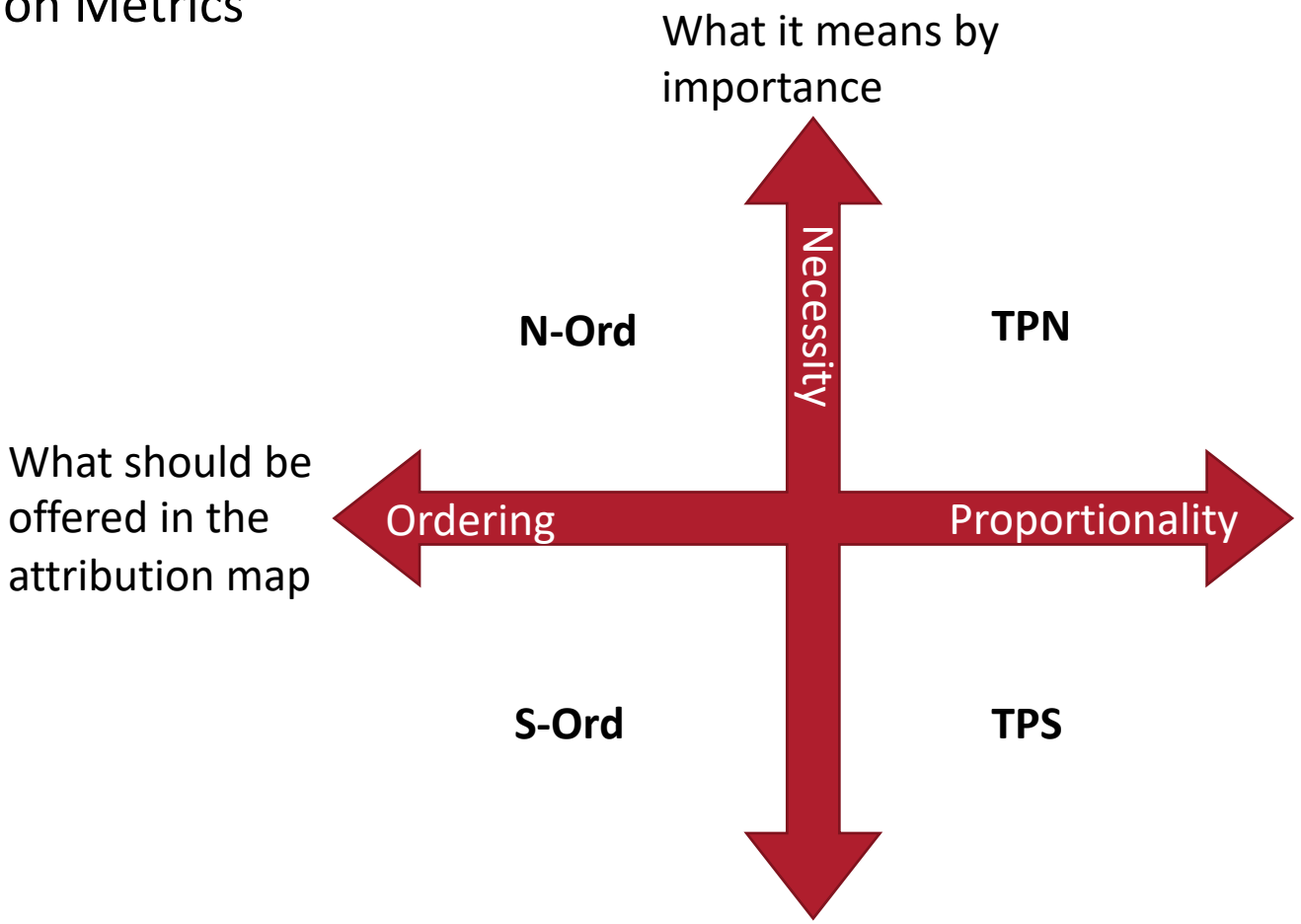
Area between the curves measures disproportionality.

Smaller area between the curves, better the Necessity/Sufficiency



Methods

Summary of Evaluation Metrics



Results

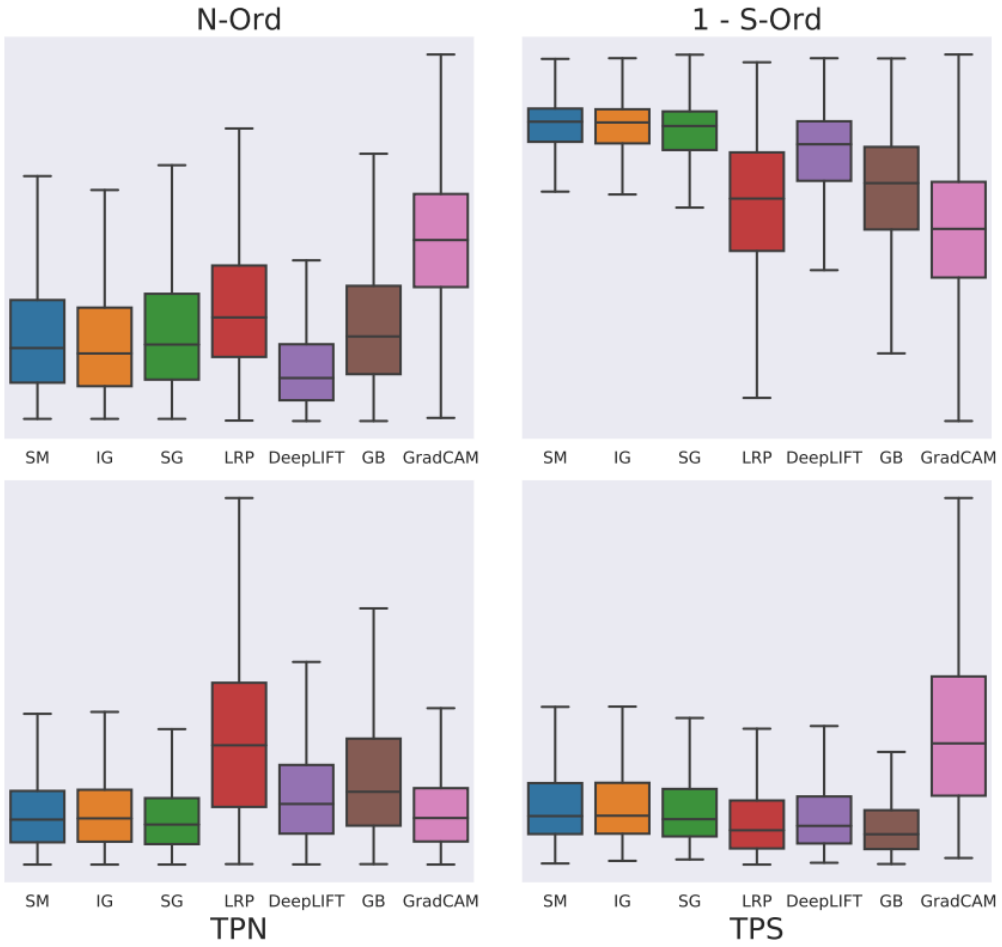
Evaluation performed with ImageNet and pretrainedVGG-16 model

Recommended Methods **under Ordering Criteria**:

- Necessity: DeepLIFT
- Sufficiency: GradCAM, LRP

Recommended Methods **under Proportionality Criteria:**

- Necessity: SmoothGrad, Saliency Map, Integrated Gradient
- Sufficiency: Guided Backpropagation, LRP, DeepLIFT

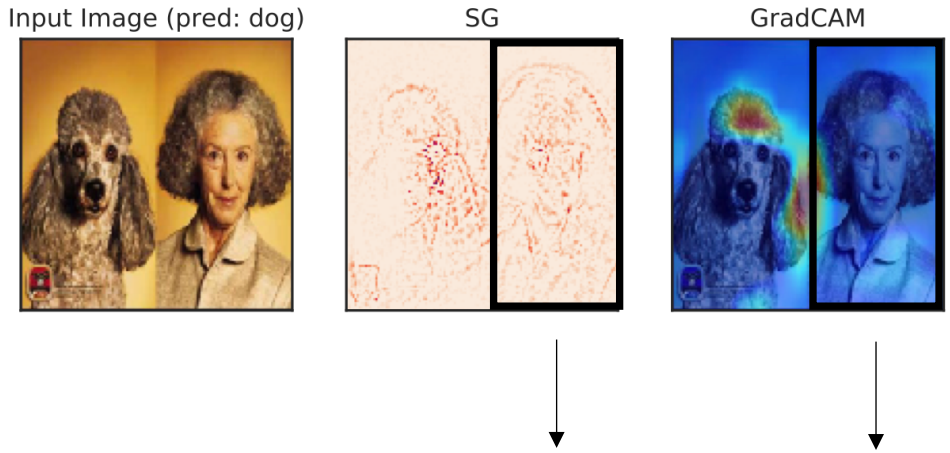


lower scores indicate the better performance

Interpret Interpretations

How do we use necessity/sufficiency and ordering/proportionality to interpret different attribution maps.

Conclusions of experiments let us impart additional interpretation to these results



The lady **IS or IS NOT** used by the model to predict “dog” class

GradCAM highlights **an area of sufficient features (Good S-Ord),** the model can make correct prediction without the lady. However, among the sufficient features, **higher scores do not strictly mean that those features are a lot more sufficient than features with lower scores (Poor TPS).**

SmoothGrad **highlights necessary features (Poor S-Ord and TPS)** , and **attribution scores are proportional to actual necessity (Good TPN).** Point clouds around the lady are less intensive compared to the dog; therefore**, the lady is less necessary compared to the dog in the prediction towards “dog” class**.

IEEE CVPR2020 WORKSHOP ON
FAIR, DATA EFFICIENT AND TRUSTED COMPUTER VISION

Thanks for watching our presentation

Contact us: zifan.wang@sv.cmu.edu

Zifan Wang, Piotr Mardziel, Anupam Datta, Matt Fredrikson

Accountable System Lab (https://fairlyaccountable.org/)

Carnegie Mellon University