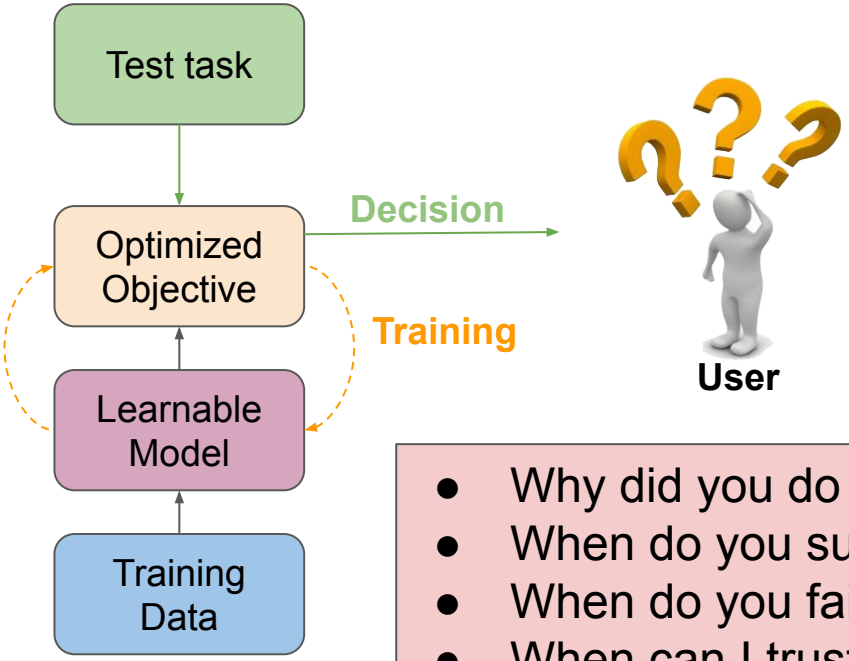
Iterative and Adaptive Sampling with Spatial Attention for Black-Box Model Explanations

Bhavan Vasu Chengjiang Long

Kitware Inc,
Clifton Park, NY, USA 12065.
{bhavan.vasu, chengjiang.long}@kitware.com

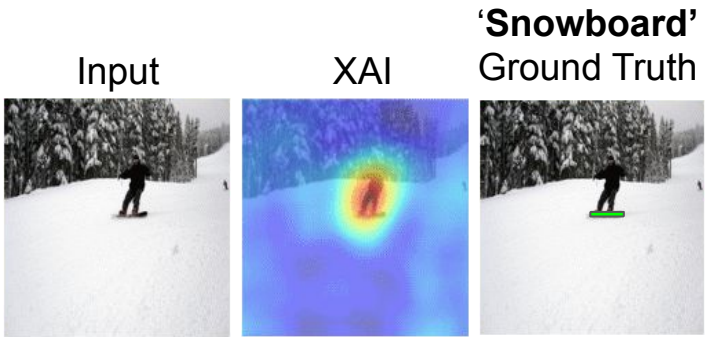
Explainable AI - Overview



User

- Why did you do that?
- When do you succeed?
- When do you fail?
- When can I trust you?

Ambiguity in XAI

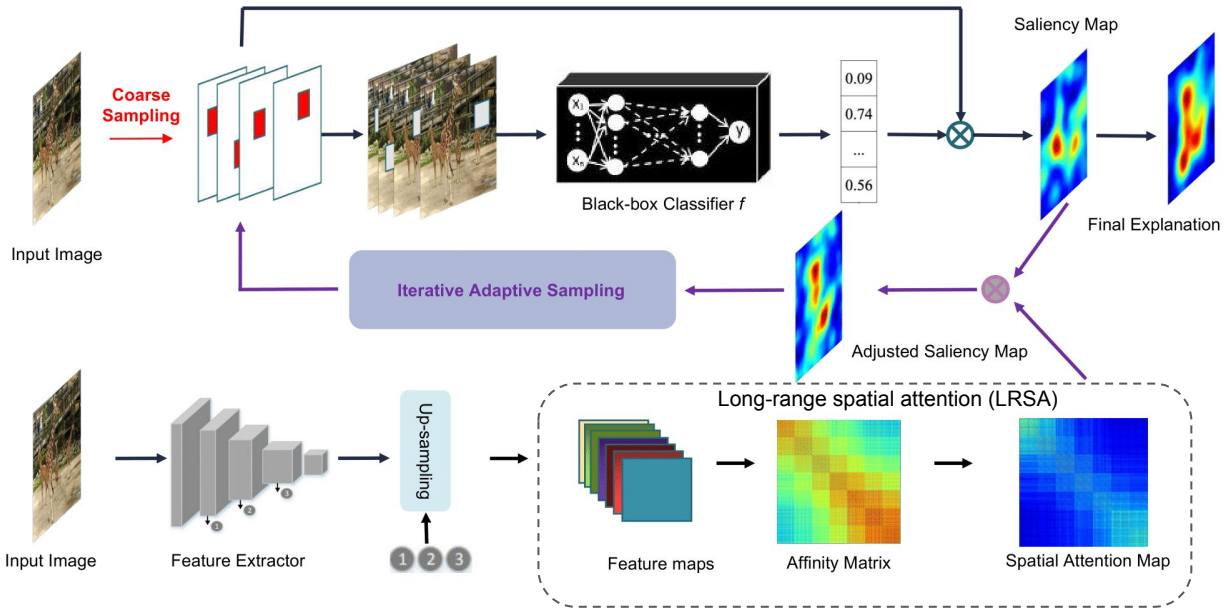


***User**: "Are the legs important?!"*

"What the model thinks as important is not necessarily what the user thinks as important."

Kitware

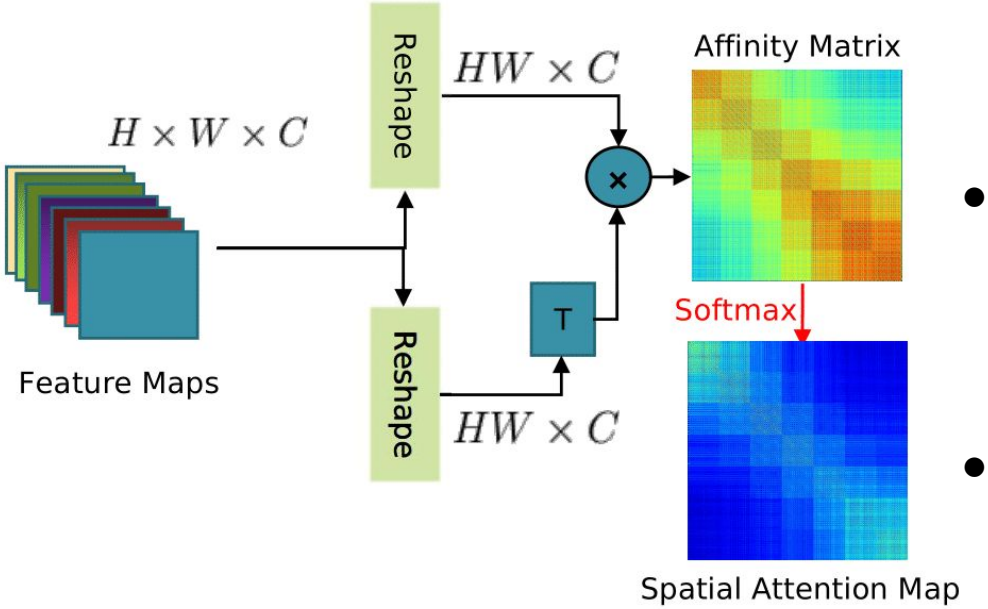
Iterative and Adaptive Sampling (IAS) - Overview



- The input image is **sampled coarsely using a sliding window** to obtain an aggregated saliency map.
- Simultaneously, we obtain a **spatial attention map** of the input image using the **LRSA module**.
- An **adjusted saliency map** is obtained after combining the **saliency map** from previous iteration and **attention map**.
- This is iteratively repeated till there is **little** or **no change** in final saliency maps.

Kitware

Long-Range Spatial Attention (LRSA) - Overview



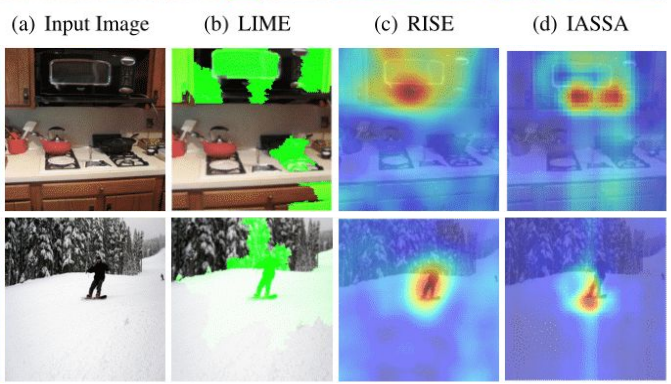
- Receptive fields limit the area of consideration to a small window in the image.
- Long-range spatial attention lets us explore and combine **long range inter-pixel dependencies** to produce an affinity matrix.
- The **output** of the LRSA module is a **spatial attention map**. Note that our LRSA module does not contain an learnable parameters.

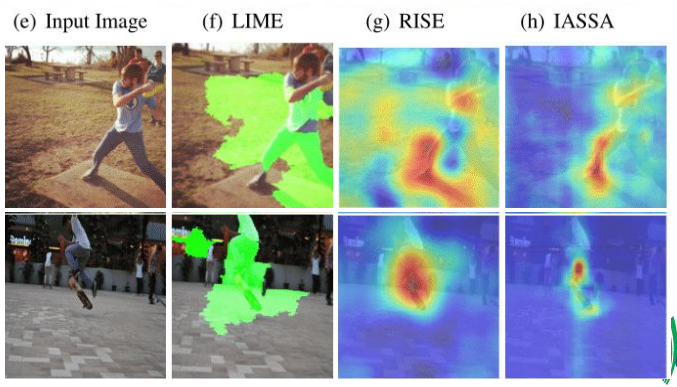
Kitware

Results comparison

Table 1. Comparative evaluation in terms of deletion (lower is better) and insertion (higher is better), F-1 (higher is better), IoU (higher is better), and Pointing Game (higher is better) scores at both image and pixel levels on the MS-COCO dataset.

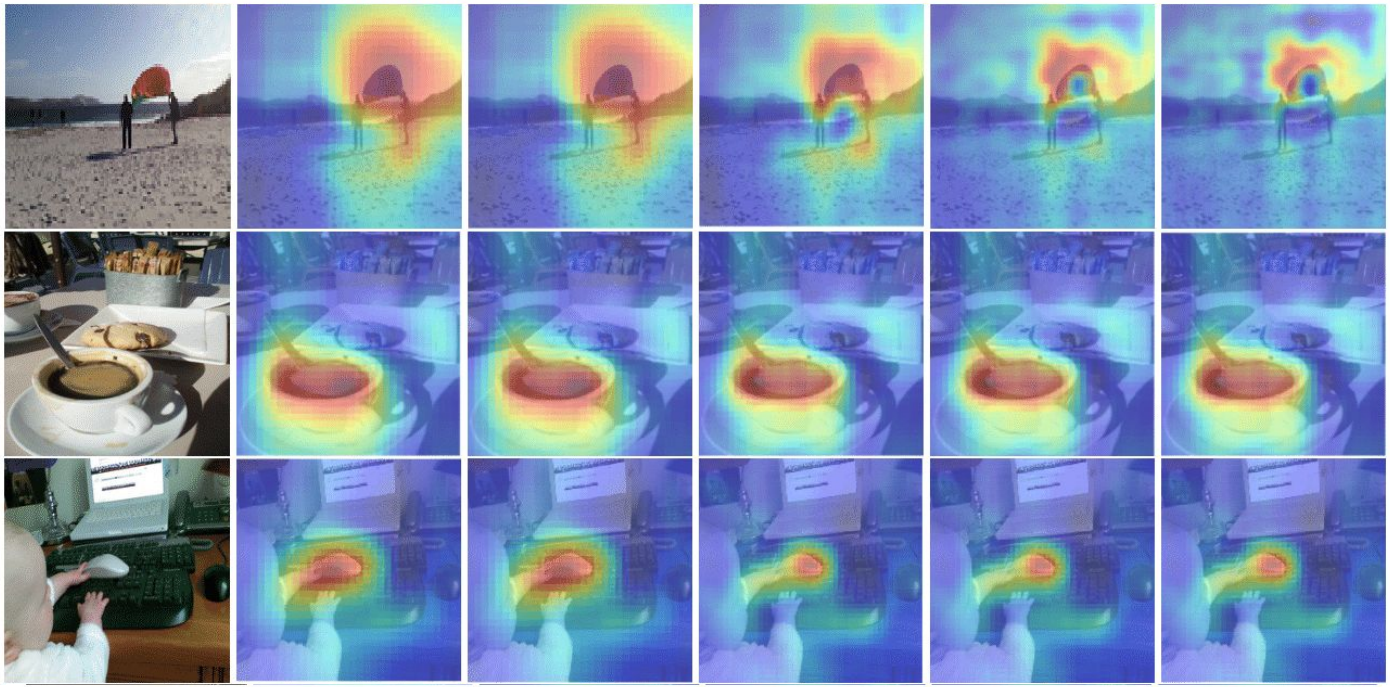
	Method	Deletion ↓	Insertion ↑	F-1 ↑	IoU ↑	Pointing Game ↑
Image-level	LIME	0.900967	0.99	0.15390	0.09745	0.16461
	RISE	**0.1847**	**1.0**	0.13837	0.13653	0.25
	IASSA	0.18803	**1.0**	**0.23658**	**0.15153**	**0.4216**
Pixel-level	LIME	10.8526e-05	10.96158e-05	1.71177e-05	1.08447e-05	0.43671e-05
	RISE	5.5423e-05	28.8669e-05	4.26672e-05	2.69240e-05	8.95937e-05
	IASSA	**5.50534e-05**	**35.33639e-05**	**10.5960e-05**	**6.9282e-05**	**17.79331e-05**





Kitware

Results across *k* iterations



(a) Input Image (b) $k = 5$ (c) $k = 10$ (d) $k = 15$ (e) $k = 20$ (f) $k = 25$

Thank you !

Please visit us at Poster 7 to know more.

Kitware