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

Source Code

Bhavan Vasu

Kitware Inc.

1712 Route 9 Suite 300, Clifton Park, NY, USA 12065

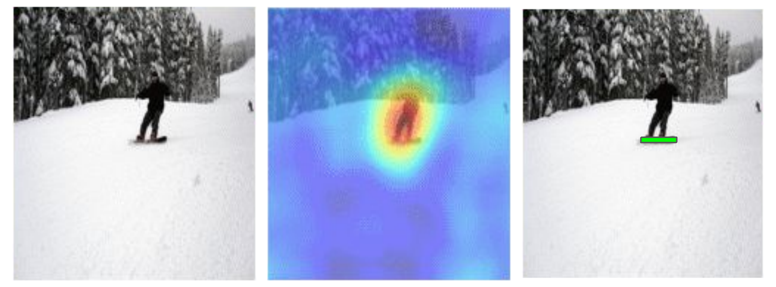
{bhavan.vasu, chengjiang.long}@kitware.com

Chengjiang Long

WACV 2020

Background & Motivation

Observation: Black-Box model explanations are generated by sampling all image regions equally to produce saliency maps. This can be computational expensive and result in coarse saliency maps due to high variance in image size.



Input XAI 'Snowboard' Ground Truth

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

Intuition: We hypothesize that sampling around important regions iteratively will result in finer saliency maps when done in a sequential manner.

Contribution: We propose a Iterative and adaptive sampler that samples around relevant regions with the help of our LRSA module. We also re-visit methods used to evaluate explanations and propose a new evaluation scheme.

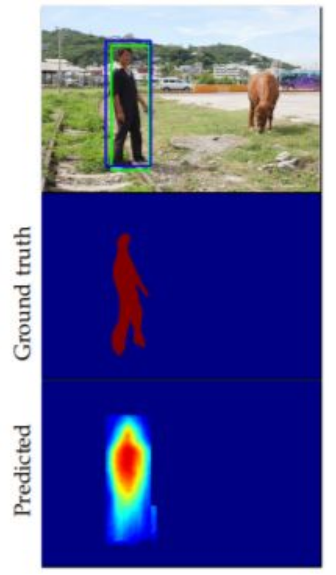
Competing Algorithms

[LIME] T. L. Pedersen and M. Benesty. lime: Local interpretable model-agnostic explanations. R Package version 0.4, 1, 2018.

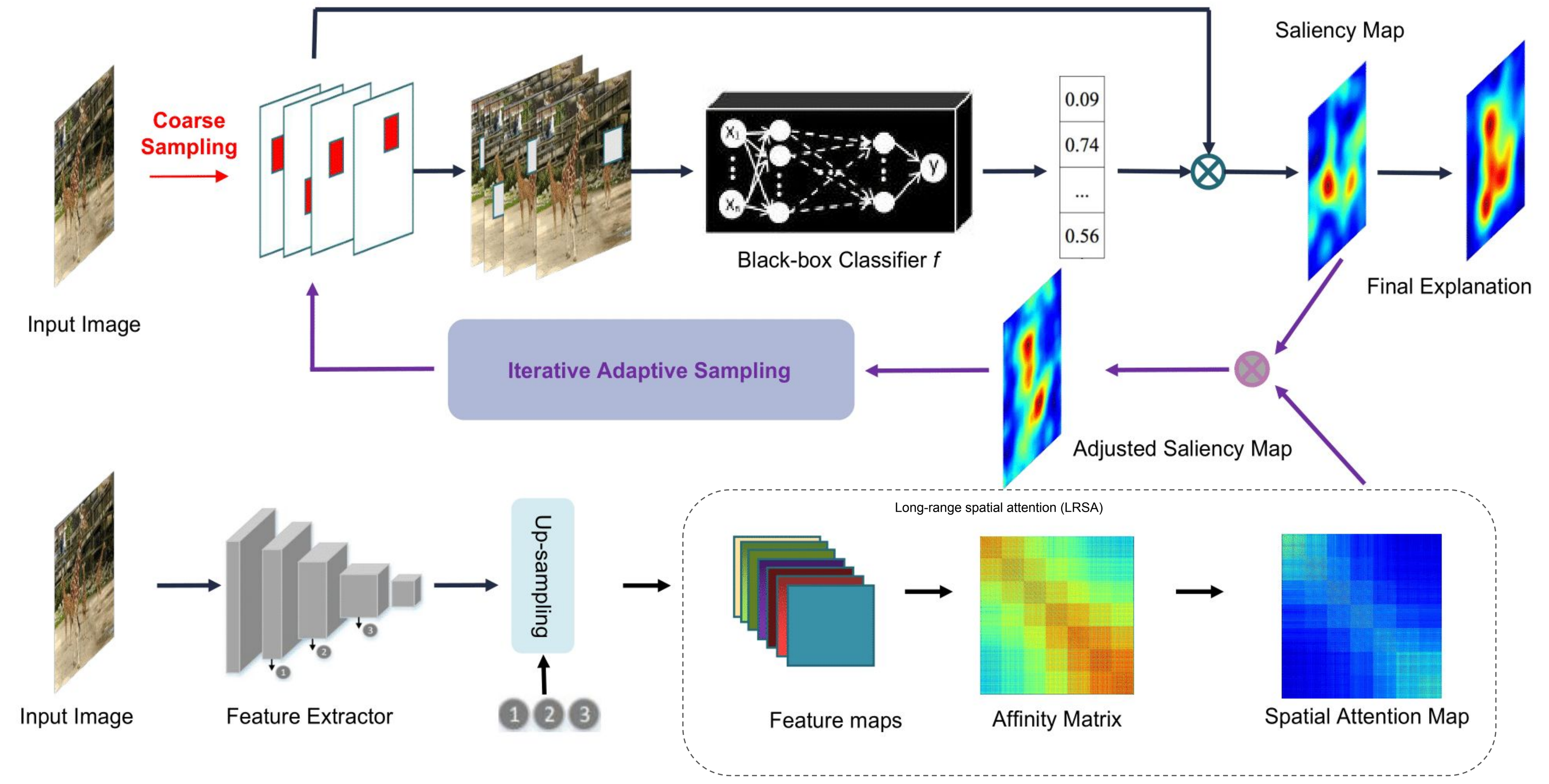
[RISE] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. BMVC, 2018.

Dataset and Metrics

- MSCOCO dataset: ~80 object categories with ~200k images.
- Evaluation metrics: Deletion, Insertion, F-1, IoU and Pointing Game.



Proposed Approach



Sampling

$$S(I,f,\lambda) = \sum_m f(I \odot M) P[M=m, M(\lambda)=1]$$

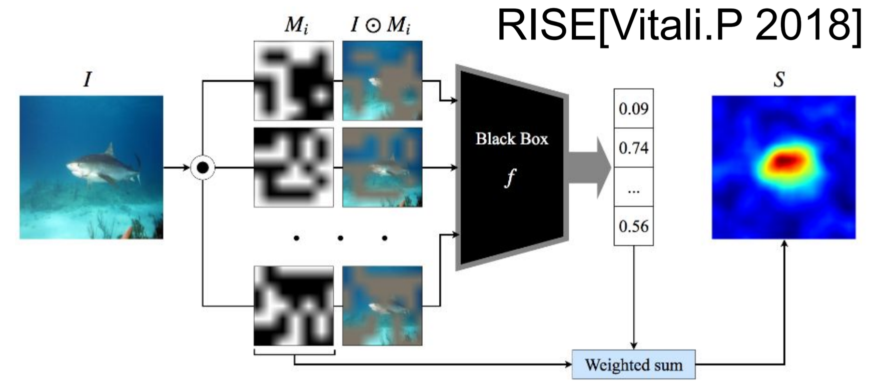
where

$$P[M=m, M(\lambda)=1] = \begin{cases} 0, & \text{if } m(\lambda)=0 \\ P[M=m], & \text{if } m(\lambda)=1 \end{cases}$$

$$S(I,f,\lambda) \approx \frac{1}{\mathbb{E}[M]\cdot N} \sum_{i=1}^{N} f(I \odot M_i).M_i(\lambda)$$

$$S'_k = \lambda S_k + (\lambda - 1)A \times S_k$$

$$\mathbf{M}_{k+1} = \mathrm{HAR}(S'_k)$$

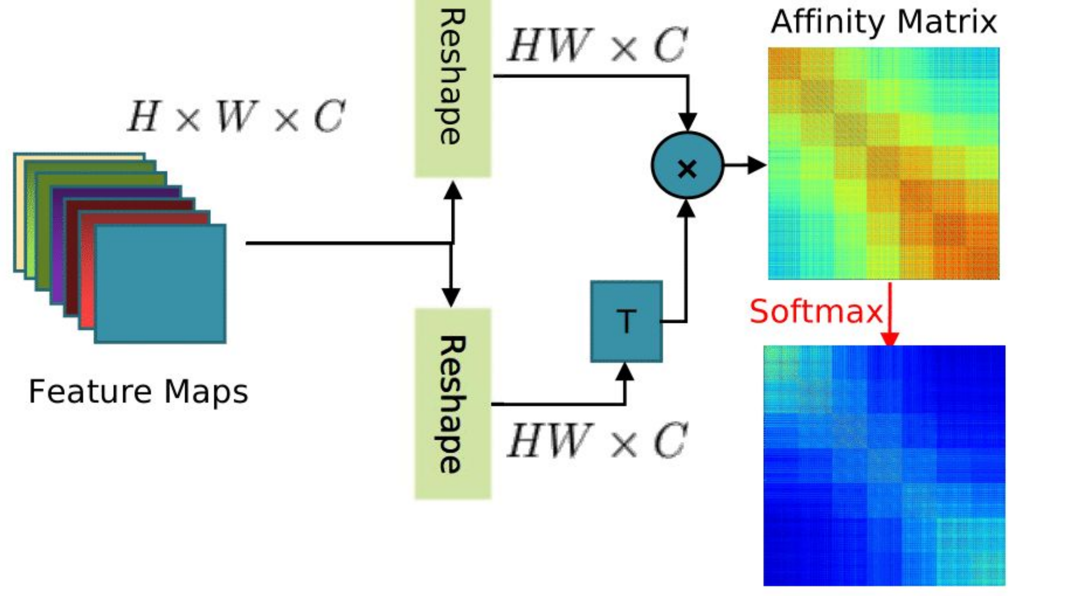


RISE[Vitali.P 2018]

Long-range spatial attention

$$\Phi(I) = \Phi_1(I) \oplus (\Phi_2(I)_\uparrow \oplus \Phi_3(I)_\uparrow \oplus \Phi_4(I)_\uparrow$$

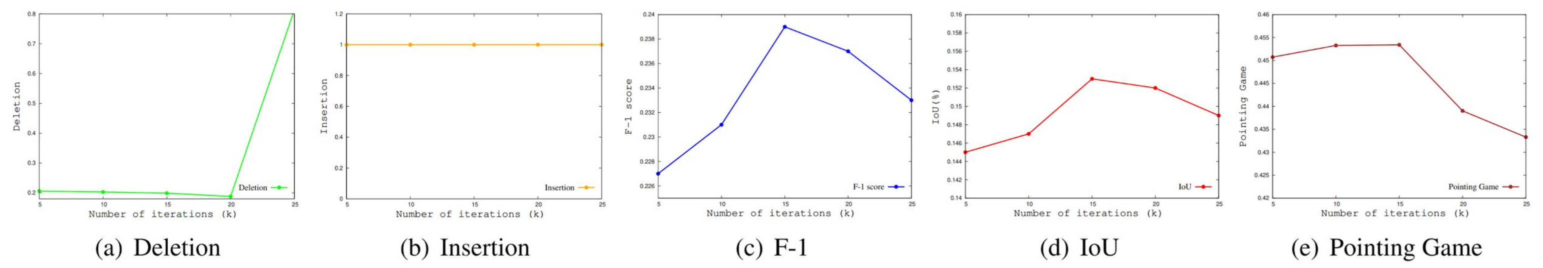
$$A = softmax(\Phi'(I) \odot (\Phi'(I))^T),$$



Experiments

Comparison with the state-of-the-art approaches

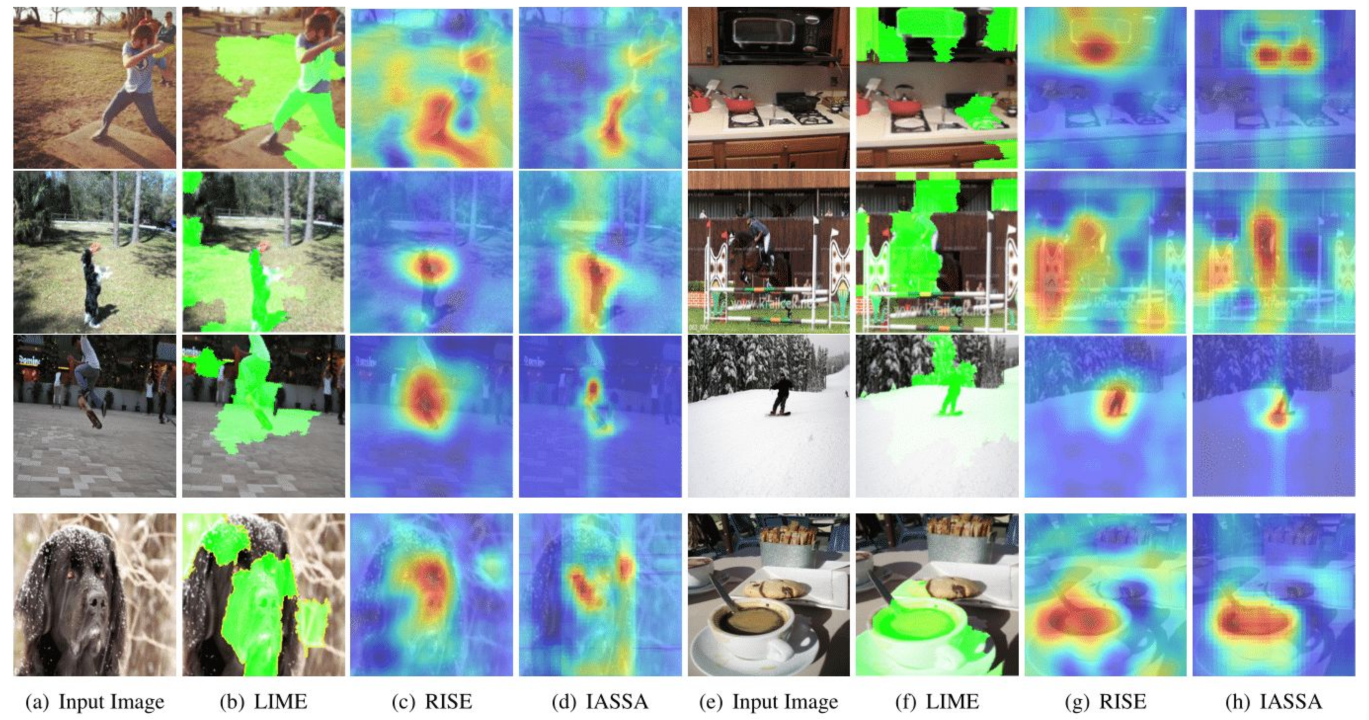
	Method	Deletion ↓	Insertion ↑	F-1 ↑	IoU ↑	Pointing Game ↑
Image-level	LIME	0.900967	0.99	0.15390	0.09745	0.16461
	RISE	**0.1847**	**1.0**	0.13837	0.13653	0.25
	IASSA	0.18803	**1.0**	**0.23658**	**0.15153**	**0.4216**
Pixel-level	LIME	10.8526e-05	10.96158e-05	1.71177e-05	1.08447e-05	0.43671e-05
	RISE	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**



(a) Deletion (b) Insertion (c) F-1 (d) IoU (e) Pointing Game

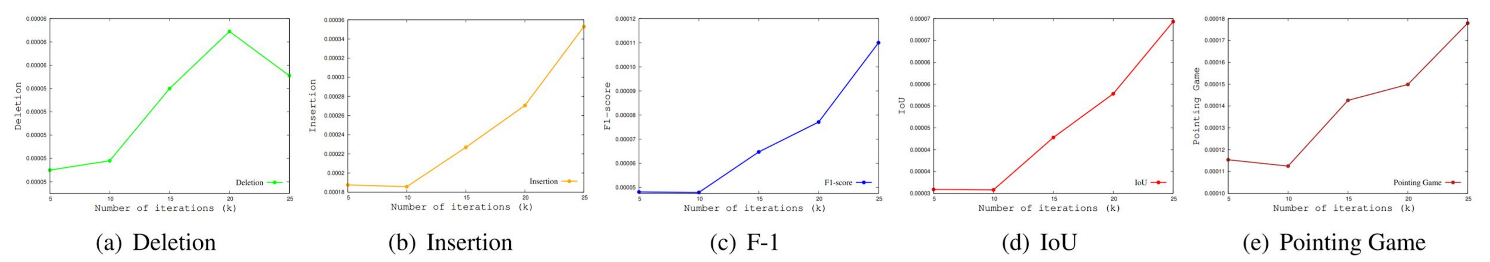
Image-level Performance vs Number of Iteration (k)

Qualitative comparison across methods



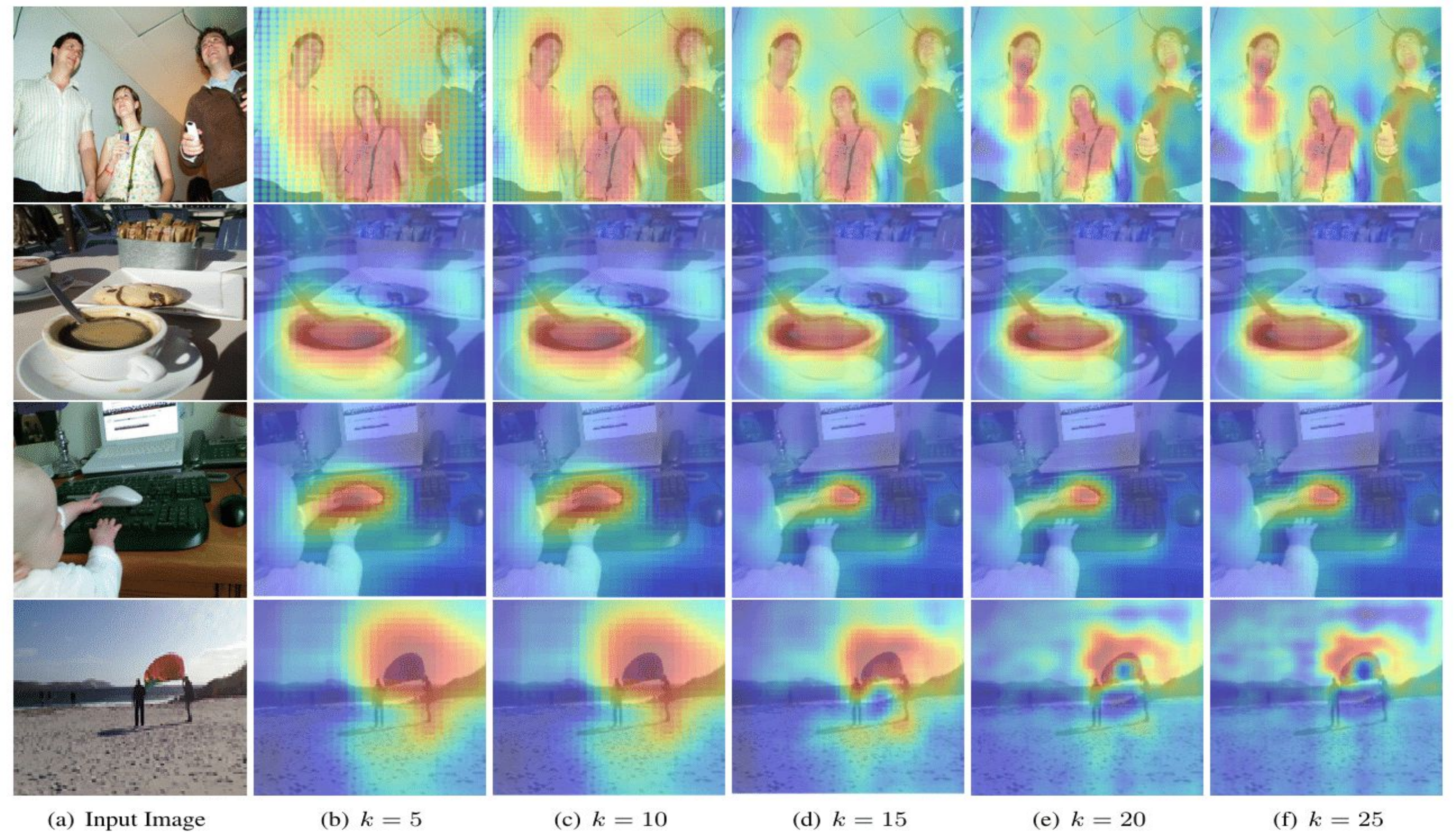
(a) Input Image (b) LIME (c) RISE (d) IASSA (e) Input Image (f) LIME (g) RISE (h) IASSA

Pixel-level Performance vs Number of Iteration (k)



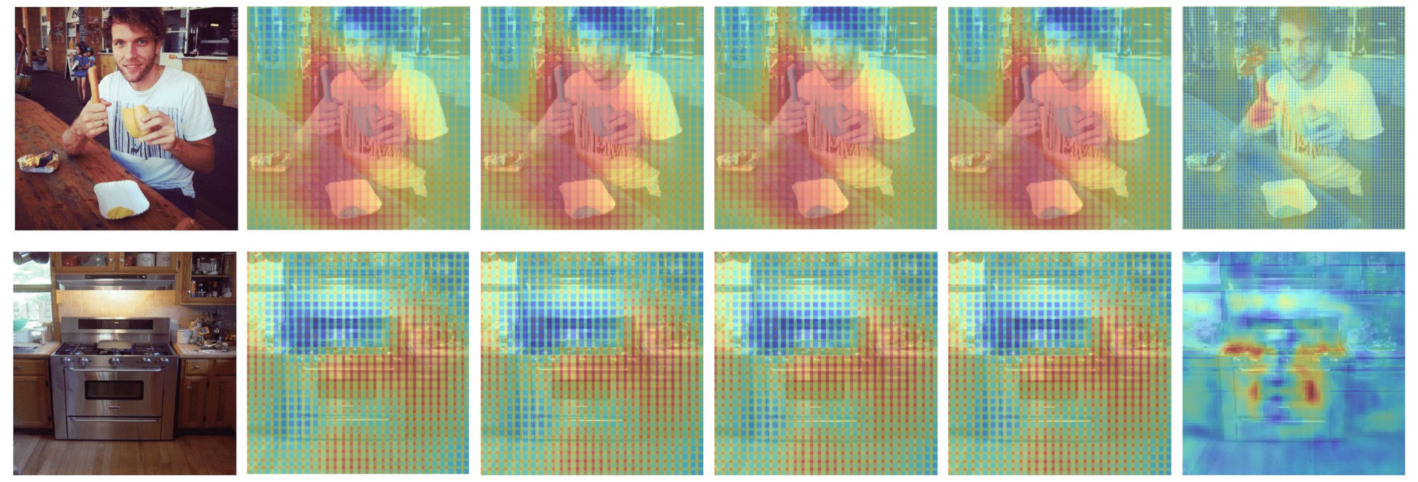
(a) Deletion (b) Insertion (c) F-1 (d) IoU (e) Pointing Game

Qualitative comparison with increasing iterations (k)



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

Sampling artifacts caused due to sliding window



Conclusion & Future Work

- We propose a novel iterative and adaptive sampling with a parameter-free long-range spatial attention for generating explanations for black-box models.
- Future work involves coming up with a universal evaluation protocol to evaluate different kinds of explanations and feed explanations agreed upon by user back into the model as 'advice'.