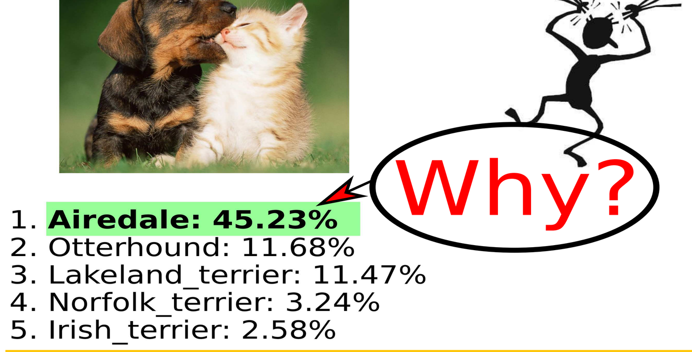


A generalizable saliency-map based interpretation of model outcome

Shailja Thakur and Sebastian Fischmeister {s7thakur, sfischme}@uwaterloo.ca

UNIVERSITY OF WATERLOO Department of Electrical and Computer Engineering

Problem

Problem1: Given a classification model, an input vector, and a confidence score for the target, weigh the input samples in the order of their importance for classification of the input to the target class.

Problem2: Given a relevance mask and a generative model, identify the distribution of acceptable variations for the important input samples.

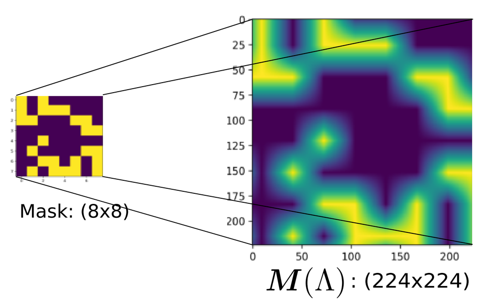
Image (church) with relevant mask | Variations for the relevant region | Complete Images

Motivation

1. Interpretability ↓ as model complexity ↑
2. Model outcome critical in decision-making
   (a) Medical applications
   (b) Self-driving cars
   (c) Safety-Critical autonomous systems
3. Risk to human lives, property, and the environment

Notations & Initializations

Input $\boldsymbol{X} : \Lambda \in \{1,...,H\} \times \{1,...,W\} \rightarrow \mathbb{R}^3$, **Output** A confidence score in the target class $c \in \mathbb{R}^C$, **Classifier** $f : \boldsymbol{X} \rightarrow \mathbb{R}^C$



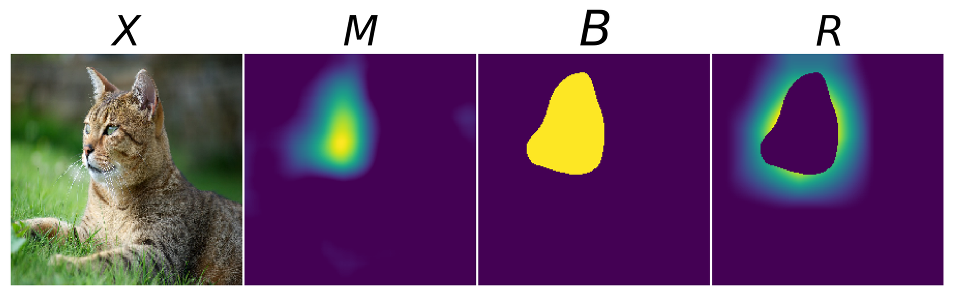
Preserved: Λ_{on}

Masked Pixels: Λ_{off}

$\Lambda = \Lambda_{\text{on}} \cup \Lambda_{\text{off}}$

$\boldsymbol{M}(\Lambda) = M(\Lambda_{\text{on}}) + M(\Lambda_{\text{off}})$

M: Estimates saliency-map, **B**: Binary bounding-box for **M** using a threshold, **R**: Inverted and convolved (kernel $(s \times s)$) version of **B**



Saliency-Map Algorithm

Key: The classifier's output is sensitive to the changes in the input.

Mask Estimation Algorithm I

Input: Input image $I \in \mathcal{I}$, Target class $c \in \mathcal{C}$
Output: saliency map M
Initialisation : $M_0 \in \mathbb{R}^{h \times w}$

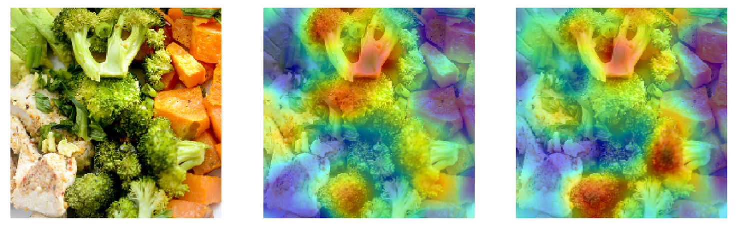
for $i = 1$ to N **do**
  $M \leftarrow$ upsample M_{i-1}
  $p_i(c|(I \odot M)) = f(I \odot M)$
  $\Delta p \leftarrow |p_i - p_{i-1}|$
  $\Lambda_1 \leftarrow$ randomly select $n_1 p_i$ pixels of Λ_{on}
  $\Lambda_2 \leftarrow$ randomly select $n_2(1 - p_i)$ pixels of Λ_{off}
  $\Lambda_{\text{on}} \leftarrow \Lambda_1 \cup \Lambda_2$
  $\Lambda_{\text{off}} \leftarrow \Lambda \setminus \Lambda_{\text{on}}$
  if $(\lambda \in \Lambda_{\text{on}})$ **then**
    $M_i(\lambda) \leftarrow M_{i-1}(\lambda)\Delta p$
  end if
  if $(\lambda \in \Lambda_{\text{off}})$ **then**
    $M_i(\lambda) \leftarrow 0$
  end if
  $V = \sum_{x,y} \|M_{i-1}^{xy+1} - M_{i-1}^{xy}\|^2 + \sum_{x,y} \|M_{i-1}^{x+1y} - M_{i-1}^{xy}\|^2$
  $M_i \leftarrow M_i + \eta V M_{i-1}$
end for
return M → **Estimated Saliency-Map**

Compute model's confidence (p_i) in target class (c) using masked input $(I \odot M)$

Revise the set of preserved (Λ_{on}) and masked pixel (Λ_{off})

Update Mask

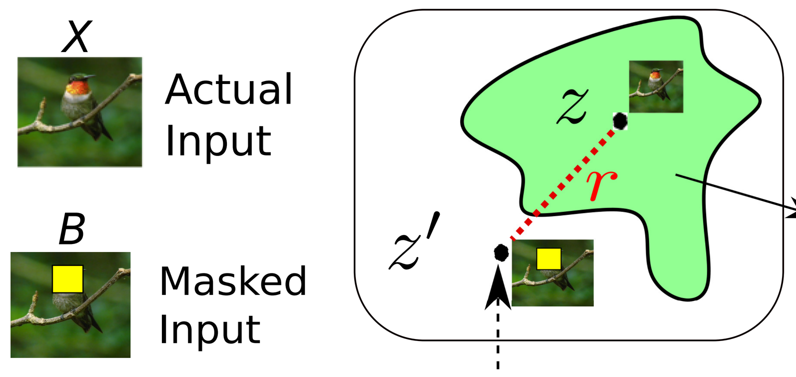
Explanation using saliency based approach lacks consistency in the detected relevent regions across runs.



Alternate Variations

Key: Latent space (z) is invariant to small preturbations (rotation, inversion, scaling, and shear).

Generative Adverserial Network



Use backpropagation, learn the encoding (z') by minimizing the loss $L(z') = L_{\text{Reconstruction}} + L_{\text{Contextual}} = r$.

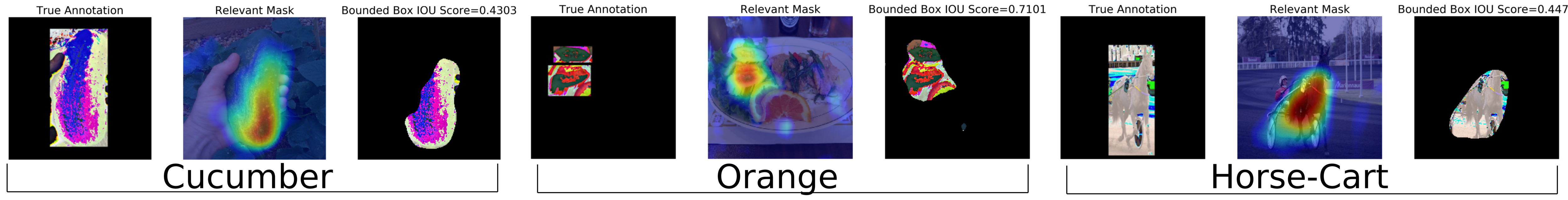
$L_{\text{Reconstruction}}$ is the MSE(x, x') and $L_{\text{Contextual}}$ determines whether x' is realistic.

Reconstructed variant of the input is,

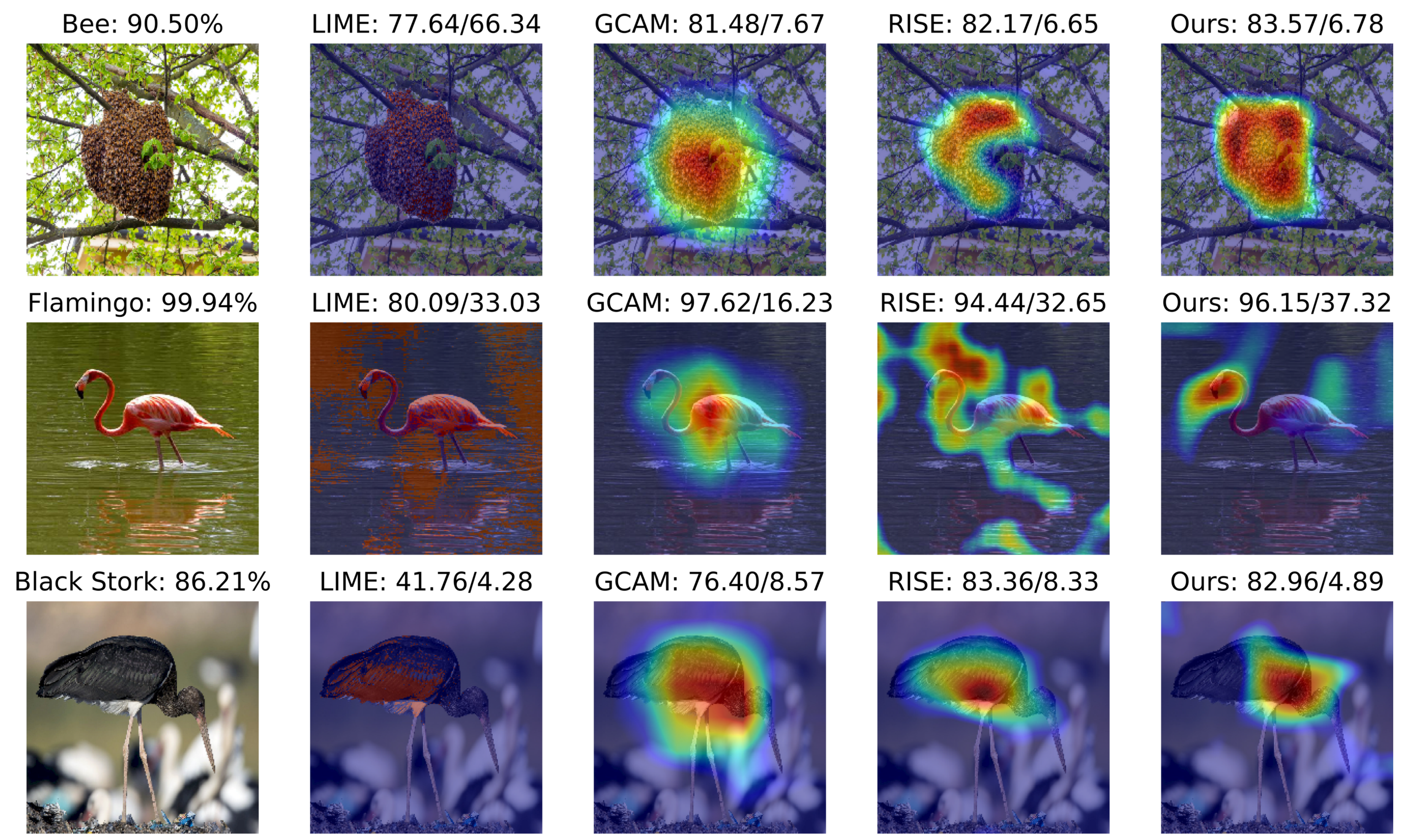
$$X' = (X \odot (\mathbf{1} - B)) + (B \odot G(z')) \tag{1}$$

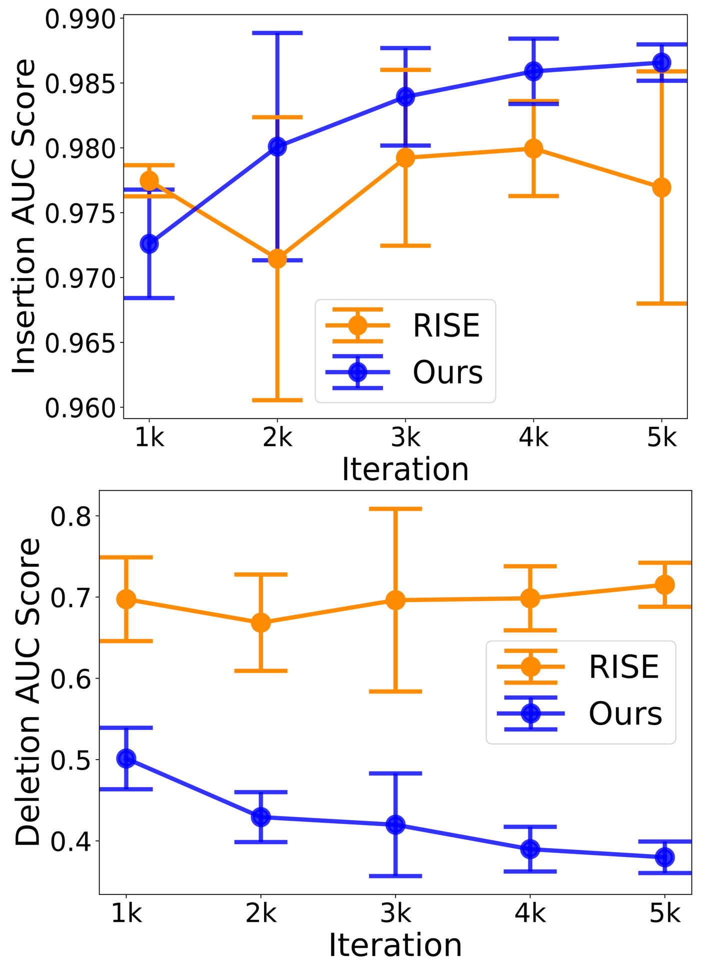
Results

IOU (Intersection Over Union) Score: Amount of overlap between saliency region and the annotated box



Insertion/Deletion Metric: This metric captures the sensitivity of the model to the insertion of the pixels from the relevant region of the input using an average AUC (Area Under the Curve) score.

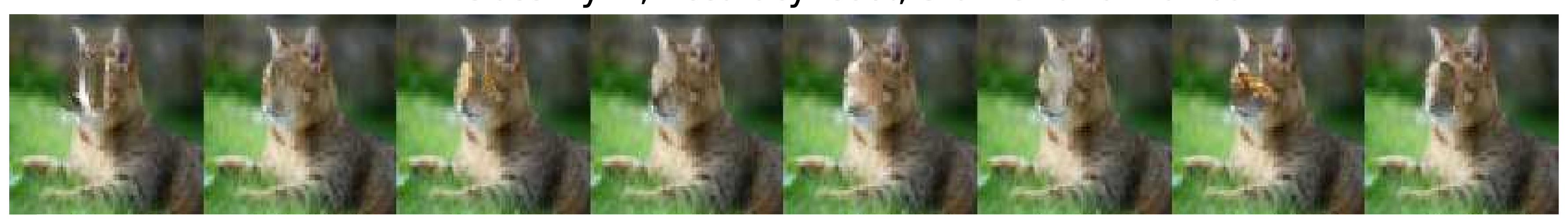




First figure, Saliency map generated for the target-specific image classification using our approach, RISE [1], GCAM [2], and LIME [3] and the AUC scores (%) of insertion/deletion metrics [1]. Second figure shows the convergence of the AUC score of insertion/deletion for the saliency map of an input image using our approach and RISE [1] over the iterations.

Acceptable variations (X′)

Class: Lynx, Accuracy: 89%, Std Deviation: 0.2%



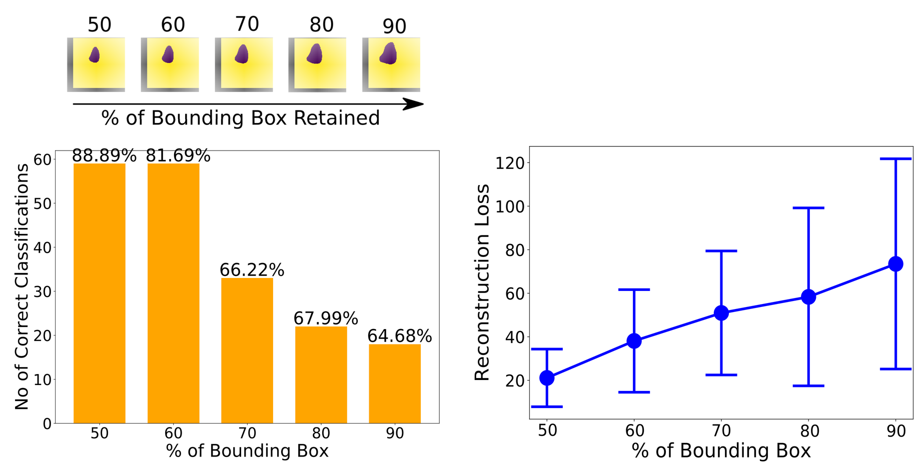


Figure (bottom left) shows the number and accuracy of correct classifications using the reconstructed images over different sizes of bounding boxes (B).

Figure (bottom right) shows the reconstruction loss over the different sizes of bounding boxes (B).

Open Questions

1. Can the set of acceptable variations be expanded to idenify adverserial examples?
2. Can the approach be applied for input of type time-series?

References

[1] V. Petsiuk, A. Das, K. Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models 2018

[2] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization 2017

[3] M. Ribeiro, S. Singh, C. Guestrin "Why Should I Trust You?": Explaining the Predictions of Any Classifier

[4] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A. Efros Context Encoders: Feature Learning by Inpainting