Black-Box Explanation of Object Detectors via Saliency Maps
Vitali Petsiuk¹, Rajiv Jain², Varun Manjunatha², Vlad I. Morariu², Ashutosh Mehra², Vicente Ordonez³, Kate Saenko¹,³
Boston University, Adobe Inc., MIT-IBM Watson AI Lab, University of Virginia

Overview.
We propose D-RISE, a method for generating visual explanations for the predictions of object detectors. Utilizing the proposed similarity metric that accounts for both localization and categorization aspects of object detection allows our method to produce saliency maps that show image areas that most affect the prediction. D-RISE can be considered “black-box” in the software testing sense, as it only needs access to the inputs and outputs of an object detector.

D-RISE
Our method D-RISE attempts to explain the detections (bounding-box+category) produced for this image by an object detector. We convert target detections that need to be explained into detection vectors d_t. We sample binary masks and run the detector on the masked images to obtain proposals. We compute pairwise similarities between targets and proposals to obtain weights for each mask. Finally, the weighted sum of masks is computed to produce saliency maps.

Similarity metric
To compute the similarity score between the target vector and the proposal vector, three components should be considered: localization (L), classification (P) and objectness (O).

Intersection over Union

Cosine similarity (Optional)

Objectness score

Evaluation
We evaluate the adapted classification saliency metrics, Pointing Game and Deletion/Insertion, to compare it against the classification-based methods. We also evaluate the ability of saliency maps help user identify which of the two models is better.

Average saliency
To explain the model from a more holistic perspective and find common patterns we compute average saliency maps for each category of MS-COCO dataset.

Conclusions
We propose a novel approach for providing saliency-based explanations for black-box object detectors. Our method is general enough to be applied to many different object detection architectures.

We demonstrate the usefulness of our method in aiding error analysis and in providing insights to model developers by means of per-class average saliency maps.

We have shown that our method is able to detect pathological biases in model behavior.

https://cs.bu.edu/vptsiuk/drise