Title

Deeply Explainable Artificial Intelligence (DExAI): High-fidelity salience for black-box networks

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
Boston University: Kate Saenko, Vitali Petsiuk

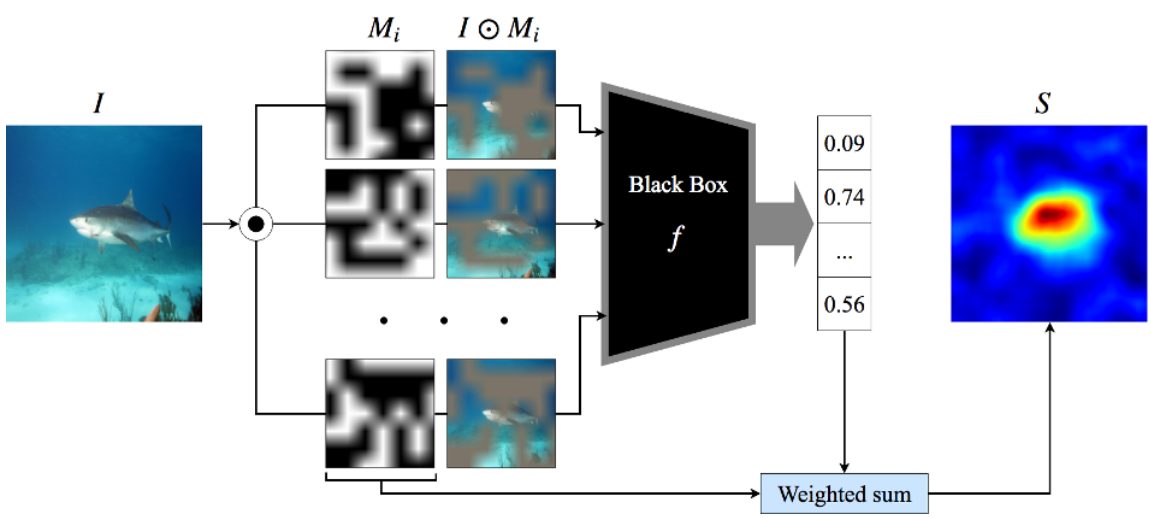
Problem Statement

This effort aims to address the problem of explainable AI for deep neural networks that take images as input, and output a class probability. The approach, called RISE, generates an importance map indicating how salient each pixel is for the model's prediction. In contrast to white-box approaches that estimate pixel importance using gradients or other internal network state, RISE works on black-box models. It estimates importance empirically by probing the model with randomly masked versions of the input image and obtaining the corresponding outputs.

Approach and Technique

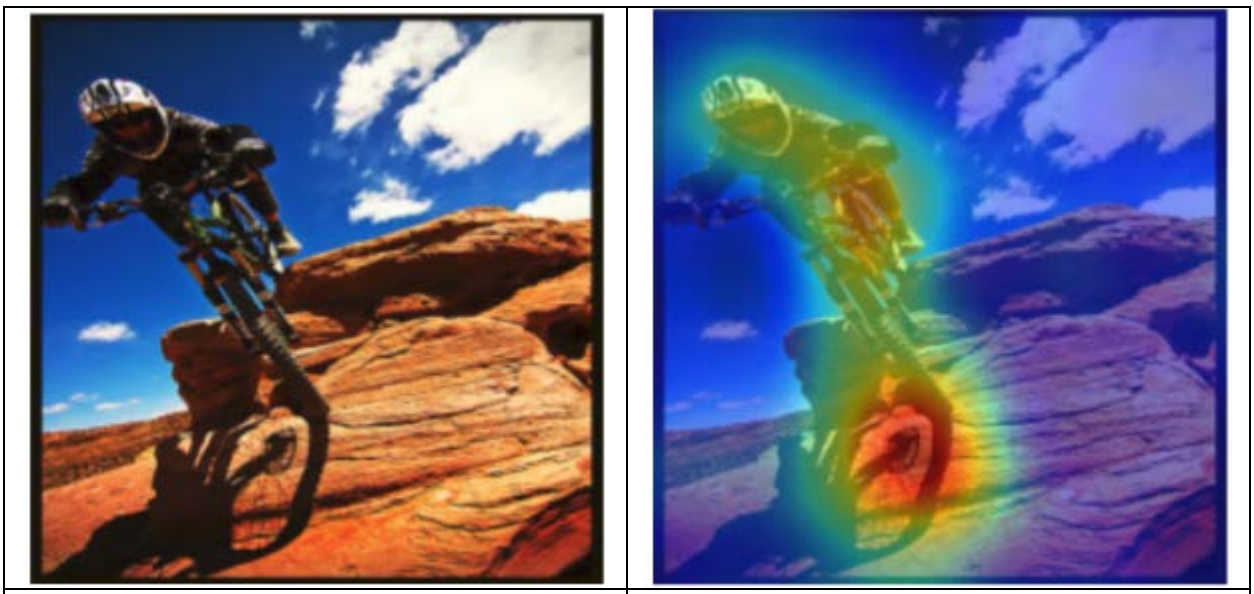
The key idea is to probe the base model by sub-sampling the input image via random masks and recording its response to each of the masked images. The final importance map is generated as a linear combination of the random binary masks where the combination weights come from the output probabilities predicted by the base model on the masked images. Figure 1 illustrates the RISE overview concept and Figure 2 illustrates an example RISE explanation.

Figure 1. RISE Overview



Approved for Public Release, Distribution Unlimited

Figure 2. Example RISE Explanation



Prediction: “Mountain bike”, 98%	Explanation for “Mountain bike” class

Key Objectives and Current Status

The objective of the effort was to generate high-fidelity saliency maps, that would highlight the regions that were actually used by the model in making its decision. The method outperformed existing approaches in terms of the metrics described below.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Explanation goodness: Deletion, Insertion	The fidelity of a saliency map explanation to the actual decision process performed by the underlying neural network	The method has been proved to have higher fidelity, i.e., better evidence grounding than previous works

Test and Evaluation

Two automatic evaluation metrics are proposed: deletion and insertion. The intuition behind the deletion metric is that the removal of the ‘cause’ will force the base model to change its decision. Specifically, this metric measures a decrease in the probability of the predicted class as more and more important pixels are removed, where the importance is obtained from the importance map. A sharp drop, and thus a low area under the probability curve (as a function of the fraction of removed pixels), means a good explanation. The insertion metric, on the other hand, takes a complementary approach. It measures the increase in probability as more and more pixels are introduced, with higher area under the curve indicative of a better explanation. Table 1 (continued) below, summarizes the results.

Approved for Public Release, Distribution Unlimited

Table 2 (continued). Measures, Objectives, and Actuals to Date

	Numerical Value		Hypothesis Results	Goodness Measure
	With Explanation	**Without Explanation**		
Explanation Goodness	0.1232/0.6766 (Grad-CAM) 0.1421/0.6618 (Sliding window) 0.1217/0.6940 (LIME) **0.1076/0.7267 (RISE)**	N/A	N/A	Deletion (lower is better) Insertion (higher is better)

Vitali Petsiuk, Abir Das, and Kate Saenko. *RISE: Randomized Input Sampling for Explanation of Black-box Models*, 2018.

For additonal information:

http://cs-people.bu.edu/vpetsiuk/rise/demo.html

Approved for Public Release, Distribution Unlimited