DANCE: Enhancing saliency maps using decoys

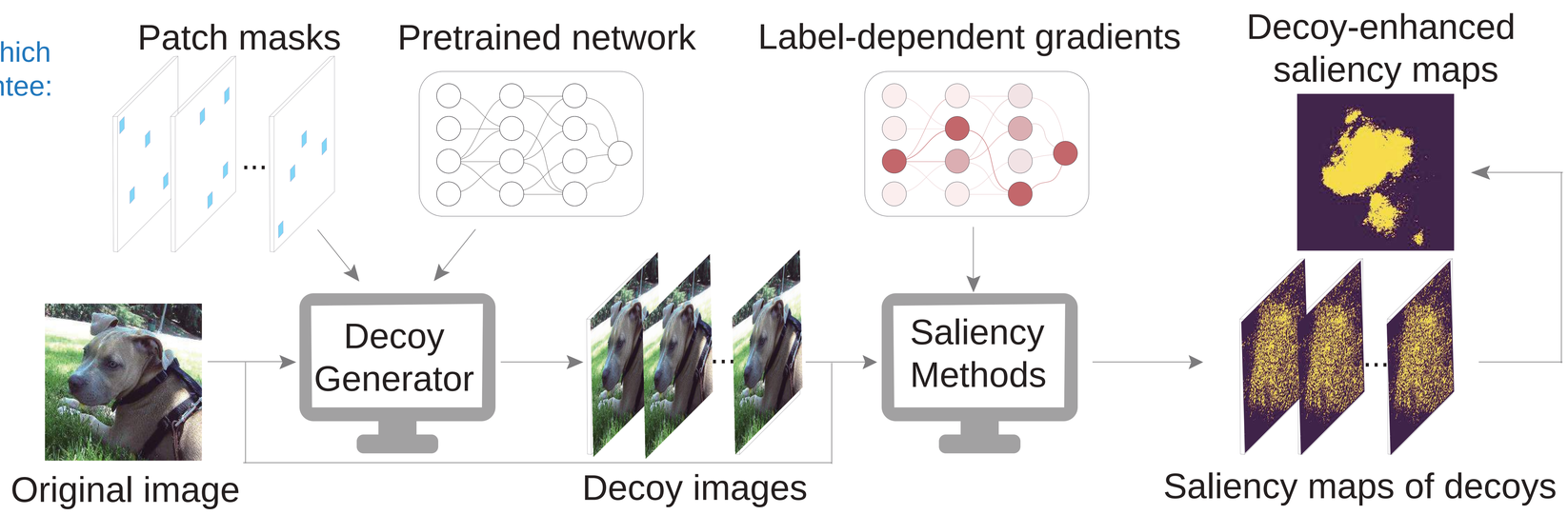
Yang Lu,[1,*] Wenbo Guo,[2,*] Xinyu Xing,[2] William Staffold Noble[1,3]

* Equal contribution
1. Department of Genome Sciences, University of Washington, Seattle, WA
2. College of Information Sciences and Technology, The Pennsylvania State University, State College, PA
3. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA

Overview

DANCE (Decoy enhANCEd saliency) is a saliency method which aims to tackle the following limitations with theoretical guarantee:

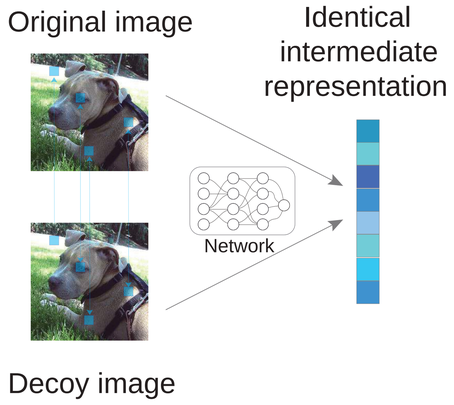
- Gradient saturation
  - Strong joint evidences together with other neighbor pixels.
  - Weak marginal evidence for individual pixel.
- Importance isolation
  - The gradient is calculated by fixing other features.
  - Smoothness in input doesn't hold in saliency maps.
- Perturbation sensitivity
  - Imperceivable noises can drastically change the saliency maps.



Decoy definition

Decoy is a perturbed variant of the input sample which is designed to preserve the the neural network's intermediate representation.

- Decoy aims to model the input variation from sensor noise or adversarial attacks.
- Both input and decoys are indistinguishable to the model prediction.
- Decoy cannot be out-of-distribution.
- Decoy can be efficiently constructed by solving an optimization problem.



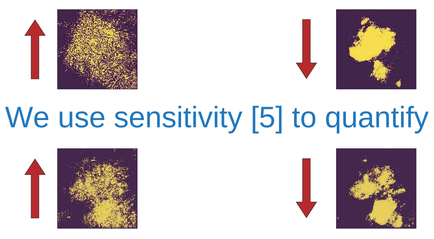
Decoy aggregation

DANCE, as well as other saliency methods, can be summarized into a "variations-and-aggregation" paradigm.

Method	Input →	Input variants →	Saliency aggregation →	Output
Vanilla Gradient [1]			NA	
Smooth Gradient [2]		+	Average	
Integrated Gradient [3]			Weighted sum	
DANCE			Pixel-wise range	

Evaluation metric

- We use fidelity [4] to quantify the coherence to the prediction



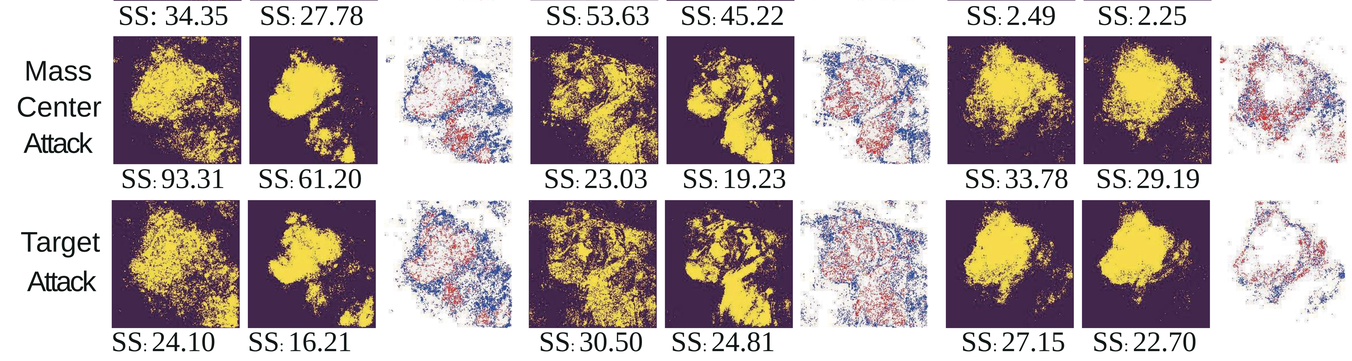
- We use sensitivity [5] to quantify the robustness to adversarial attack

Reference

[1] Simonyan, et al. "Deep inside convolutional networks: Visualising image classification models and saliency maps" arXiv:1312.6034 (2013)
[2] Smilkov, et al. " Smoothgrad: removing noise by adding noise." arXiv:1706.03825 (2017)
[3] Sundararajan, et al. " Axiomatic attribution for deep networks." ICML (2017)
[4] Dabkowski, et al. " Real time image saliency for black box classifiers." NeurIPS (2017)
[5] Alvarez-Melis, et al. " Towards robust interpretability with self-explaining neural networks." NeurIPS (2018)

Performance on ImageNet dataset

DANCE achieves more coherent saliency maps both qualitatively and quantitatively



DANCE is more robust to various types of adversarial attacks both qualitatively and quantitatively

A control study on DANCE

Two components in DANCE, both decoys and range-based aggregation, are essential.