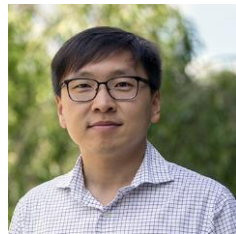


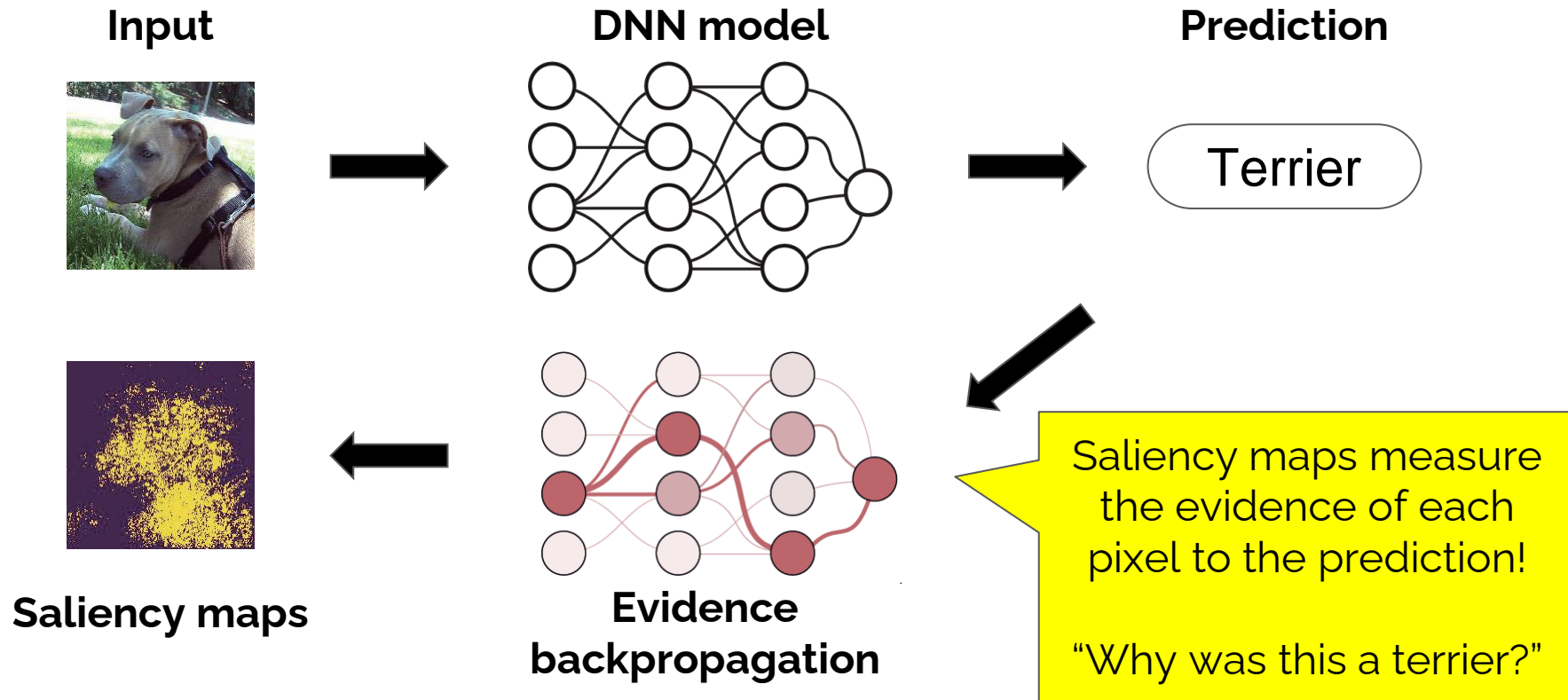


DANCE: Enhancing saliency maps using decoys

Yang Lu*, Wenbo Guo*, Xinyu Xing, William Stafford Noble



Saliency maps: the most popular interpretability method for deep neural network (DNN) models



Existing saliency maps can be summarized into a “variations-and-aggregation” paradigm

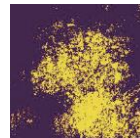
Input	Input variations	Saliency aggregation	Output
-------	------------------	----------------------	--------

Vanilla Gradient

(Simonyan et al. 2013)

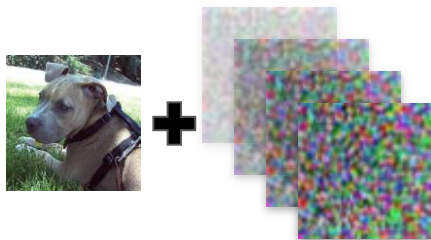


NA

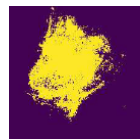


Smooth Gradient

(Smilkov et al. 2017)



Average

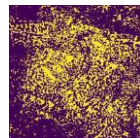


Integrated Gradient

(Sundararajan et al. 2017)



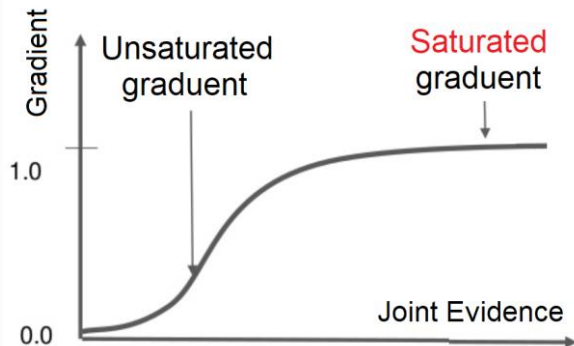
Weighted Sum



Existing saliency maps suffer from following limitations

Gradient saturation

(Shrikumar et al. 2017)



- ❑ Strong **joint** evidences together with others.
- ❑ Diminishing **marginal** evidence alone.

Isolated importance

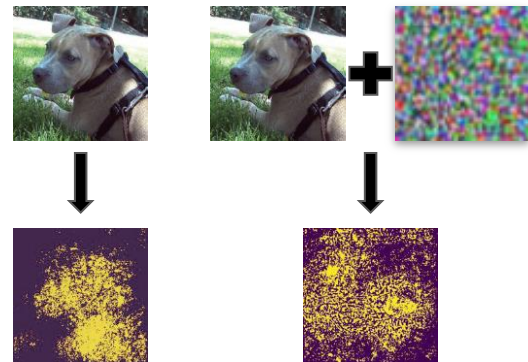
(Singla et al. 2019)



- ❑ The gradient is calculated by **fixing** other features.
- ❑ **Smoothness** in input doesn't hold in saliency maps.

Sensitive to perturbation

(Ghorbani et al. 2017)



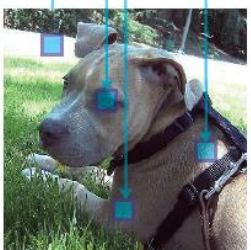
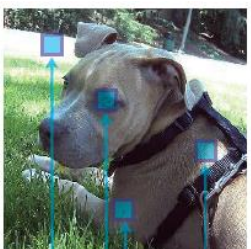
- ❑ Even **imperceivable** noises can **drastically** change the saliency maps.

We propose DANCE: decoy-enhanced saliency maps

What is a **decoy**?

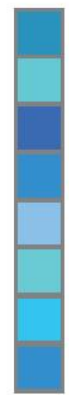
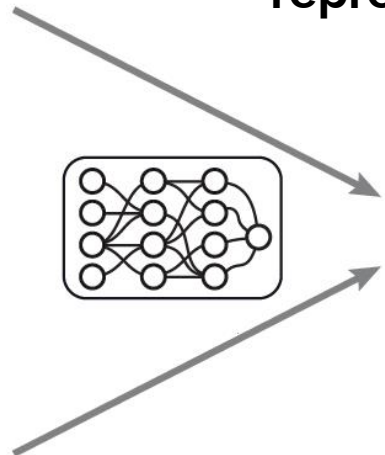


Input



Decoy

**Identical
intermediate
representation**



Decoys are constructed **independent** of the label!

Both input and decoys are **indistinguishable** to the model!

DANCE can also be decomposed into the “variations-and-aggregation” paradigm



Highlights:

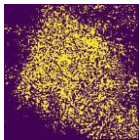
- ❑ Decoys cannot be **out-of-distribution** by design.
- ❑ Decoy can be **constructed efficiently** by optimization.
- ❑ **Theoretical soundness** in mitigating aforementioned limitations



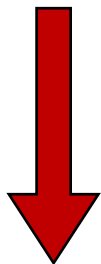
Two different metrics are used to quantitatively evaluate the performance of DANCE

Fidelity

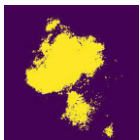
(Dabkowski & Gal, 2017)



The saliency map is **less coherent** to the prediction

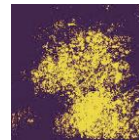


The saliency map is **more coherent** to the prediction

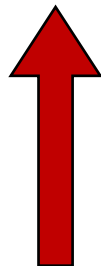


Sensitivity

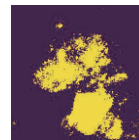
(Alvarez-Melis & Jaakkola, 2018)



The saliency map is **less robust** to adversarial attack

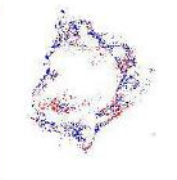
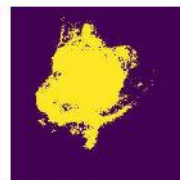
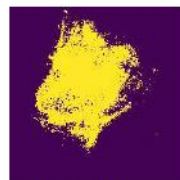
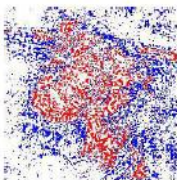
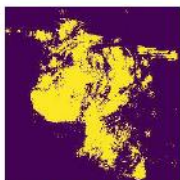
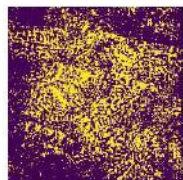
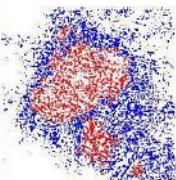
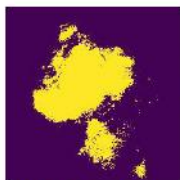
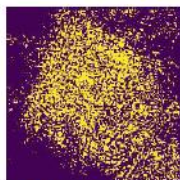


The saliency map is **more robust** to adversarial attack



DANCE achieves more coherent saliency maps both qualitatively and quantitatively

Gradient w/o decoy Gradient w/ decoys Gradient difference IntGrad w/o decoy IntGrad w/ decoys IntGrad difference SGrad w/o decoy SGrad w/ decoys SGrad difference



Terrier

SF: 11.95

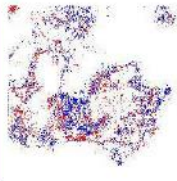
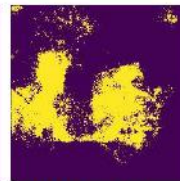
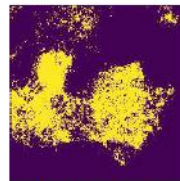
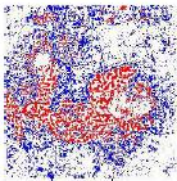
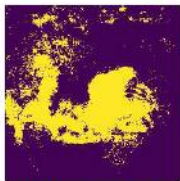
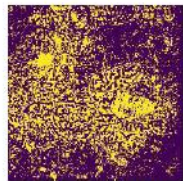
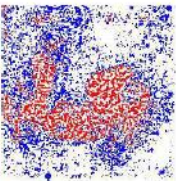
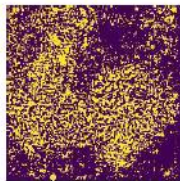
SF: 2.44

SF: 12.97

SF: 0.34

SF: 0.81

SF: 0.34



Scotter

SF: 14.17

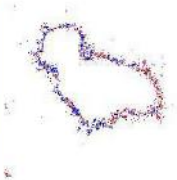
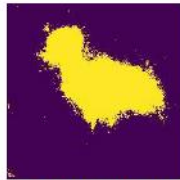
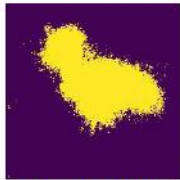
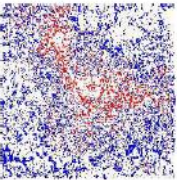
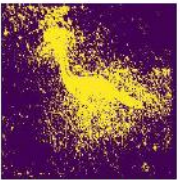
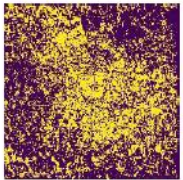
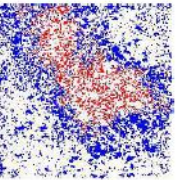
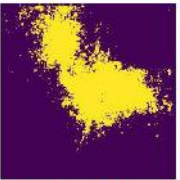
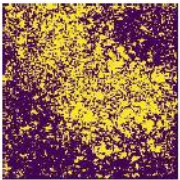
SF: 7.82

SF: 12.21

SF: 6.77

SF: 7.58

SF: 6.74



Bustard

SF: 3.05

SF: 0.08

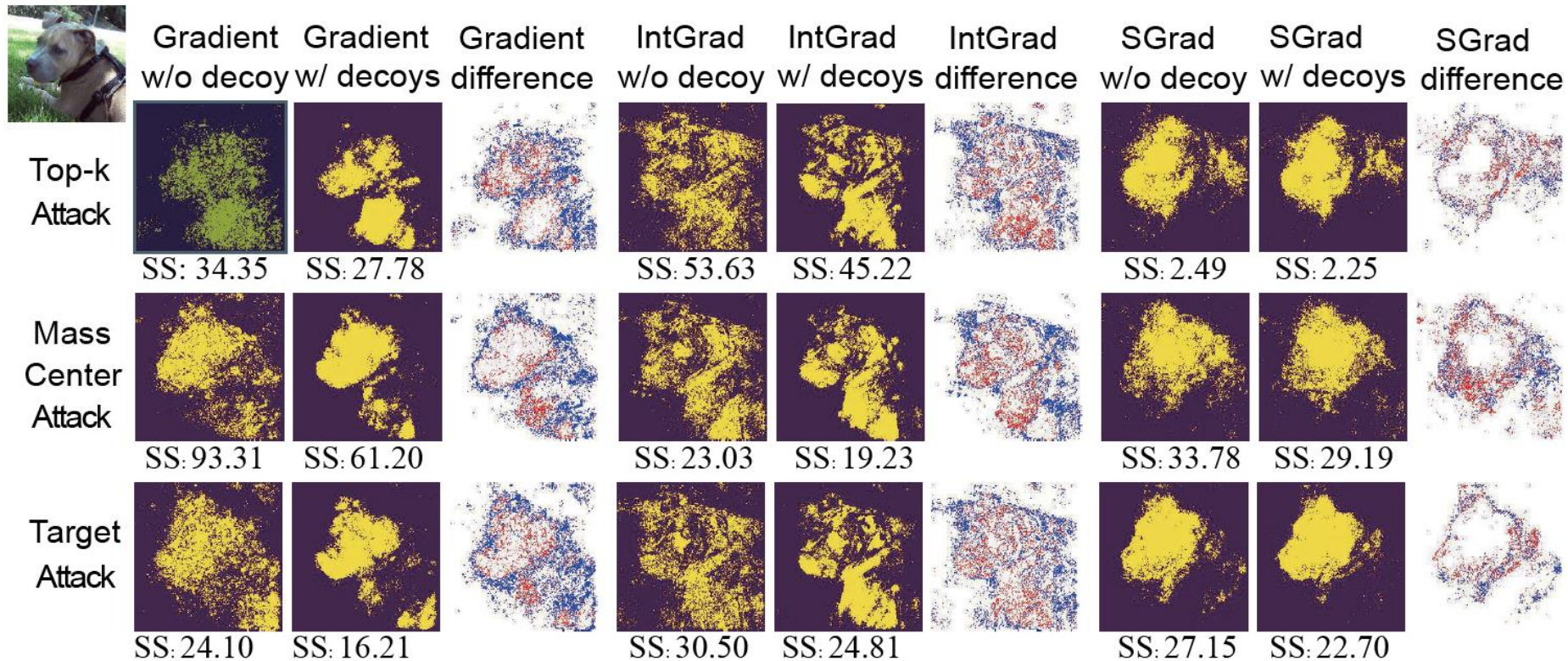
SF: 0.018

SF: 0.012

SF: 0.076

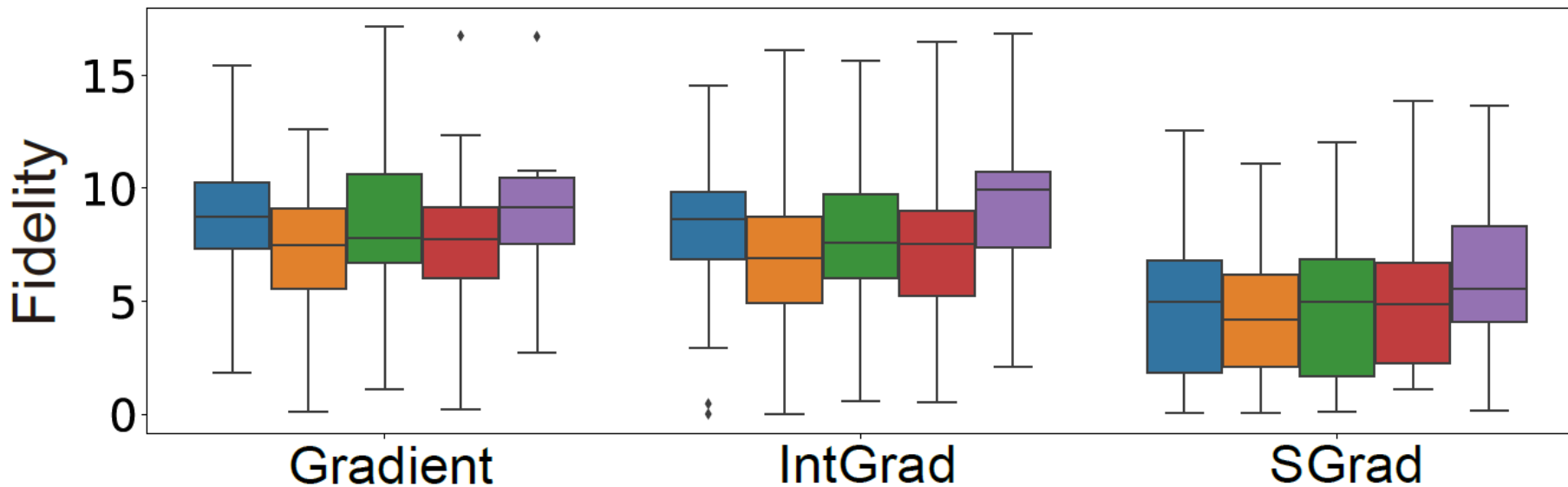
SF: 0.064

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



A control study demonstrates the necessity of both decoy and range-based aggregation steps in DANCE

Decoy variations: ■ Decoys w/ range aggregation ■ Noise w/ range aggregation
■ Without decoy ■ Constant w/ range aggregation ■ Decoys w/ mean aggregation



Conclusions

- ❑ Empirically, DANCE performs qualitatively and quantitatively better than existing methods.
- ❑ Theoretically, DANCE mitigates three limitations commonly suffered by existing methods: gradient saturation, isolated importance, and sensitivity to perturbation.
- ❑ We have demonstrated the wide applicability in a variety of domains.
- ❑ Code availability: <https://bitbucket.org/noblelab/dance>