SAM: The Sensitivity of Attribution Methods to Hyperparameters

Naman Bansal[*]

Chirag Agarwal[*]

Anh Nguyen[*]

AUBURN UNIVERSITY

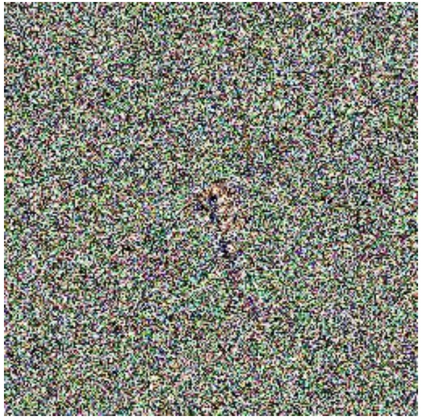
[*] Equal contribution

AIs confused by out-of-distribution examples

Nguyen, Yosinski, Clune. CVPR 2015

Goodfellow et al. 2015

Alcorn et al. CVPR 2019



cheetah 0.99

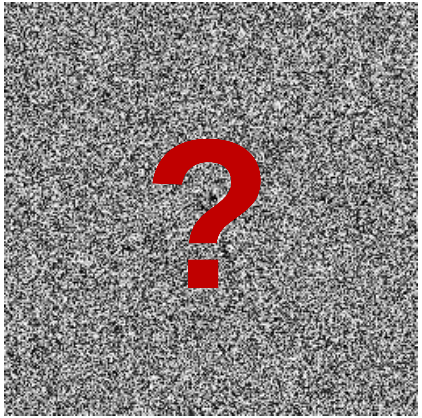
starfish 0.99

gibbon 0.99

school bus 0.98

AIs confused by out-of-distribution examples

Nguyen, Yosinski, Clune. CVPR 2015



cheetah 0.99

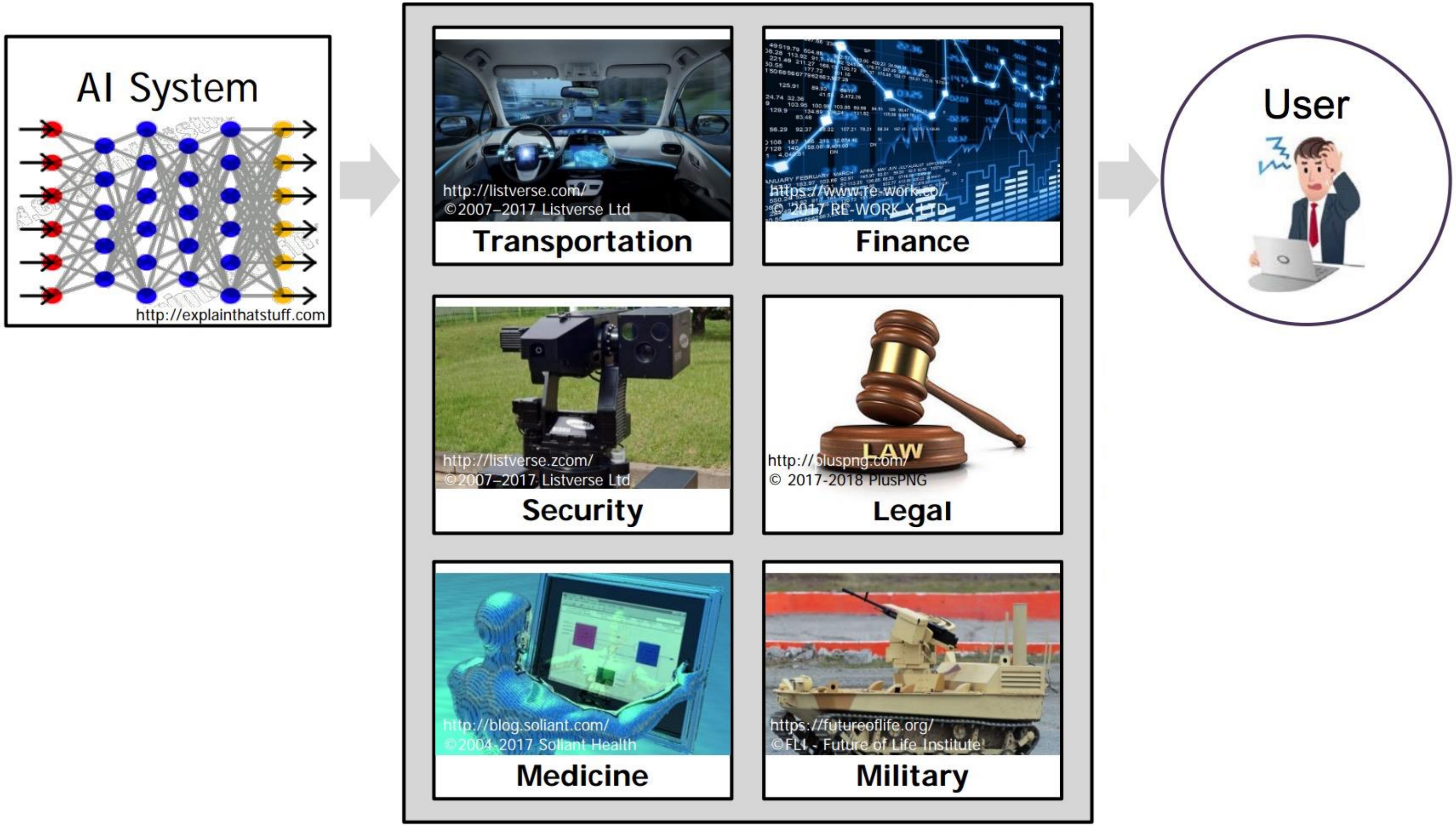
starfish 0.99

Goodfellow et al. 2015

gibbon 0.99

Alcorn et al. CVPR 2019

school bus 0.98



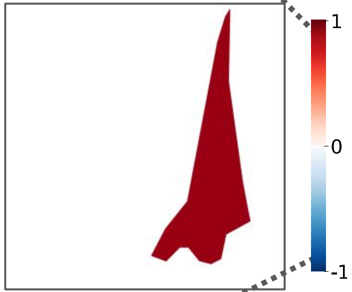
https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Attribution maps as explanations

$$x \longrightarrow \boxed{f} \longrightarrow f(x)$$

0.54 **matchstick**

for matchstick



against matchstick

attribution map
(hypothetical)

input model

$$A(x, f, \mathcal{H})$$

hyperparameters

Deconvnet: Visualizing and understanding convolutional networks. Zeiler et al. 2014

Guided-backprop: Striving for simplicity: The all convolutional net. Springenberg et al. 2015

Integrated Gradients: Axiomatic Attribution for Deep Networks. *Sundararajan et al. 2018*

CAM: Learning Deep Features for Discriminative Localization. *Zhou et al. 2016*

LIME: Why should i trust you?: Explaining the predictions of any classifier. *Ribeiro et al. 2016*

SmoothGrad: removing noise by adding noise. *Smilkov et al. 2017*

MP: Interpretable Explanations of Black Boxes by Meaningful Perturbation. Fong et al. 2017

SHAP: A Unified Approach to Interpreting Model Predictions. *Lundberg et al. 2017*

PDA: Visualizing deep neural network decisions: Prediction difference analysis. Zintgraf et al. 2017

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Selvaraju et al. 2017

Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. Chattopadhyay et al. 2017

LRP: Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation

DeepLIFT: Learning important features through propagating activation differences. Shrikumar et al. 2017

RISE: Randomized Input Sampling for Explanation of Black-box Models. Petsiuk et al. 2018

FIDO: Explaining image classifiers by counterfactual generation. Chang et al. 2019

Expected Gradients: Learning Explainable Models Using Attribution Priors. Erion et al. 2019

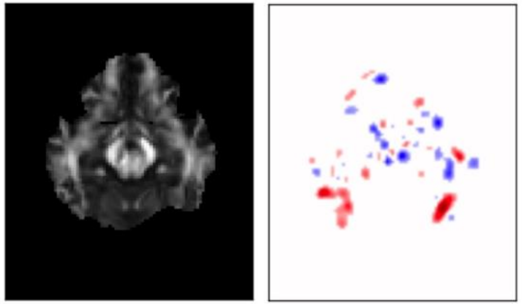
FG-Vis: Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. *Wagner et al. CVPR 2019*

Understanding Deep Networks via Extremal Perturbations and Smooth Masks. Fong et al. ICCV 2019

MP-G: Removing input features via a generative model to explain their attributions to classifier's decisions. Agarwal et al. 2020

• • •

et al. 2014
enberg et al. 20
ajan et al. 2018
2016
. Ribeiro et al. 2



Smooth
MP: Int

Natural images Fong et al. 2017

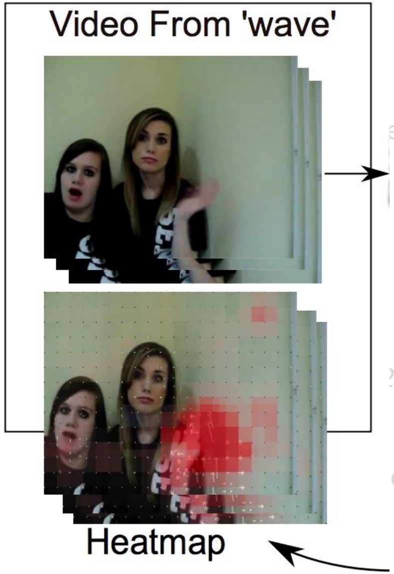
17
erturbation. Fong et al. 20

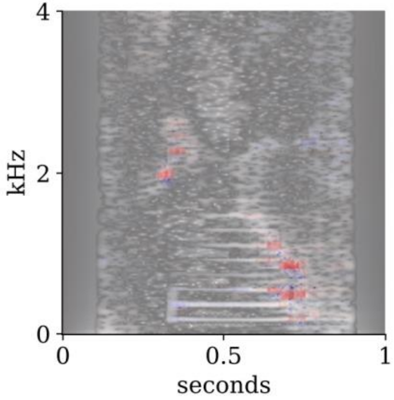
MRI brain scans Zintgraf et al. 2017

SHAP: A Unified Approach to Interpreting Model Predictions. *Lundberg et al. 2017*

PDA: Visualizing deep neural network decisions: Prediction difference analysis. Zintgraf et al. 2017

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu





FIDO: Explaini
al generation. Chang et al. 2019

Text Ribeiro et al. 2016

Expected Grad
Using Attribution Priors. Erion et

FG-Vis: Interpretable and Fine-Grained Visual Explanations for Convolutional Neur

Understanding Deep Networks via Extremal Perturba

MP-G: Removing input features via a generative mod

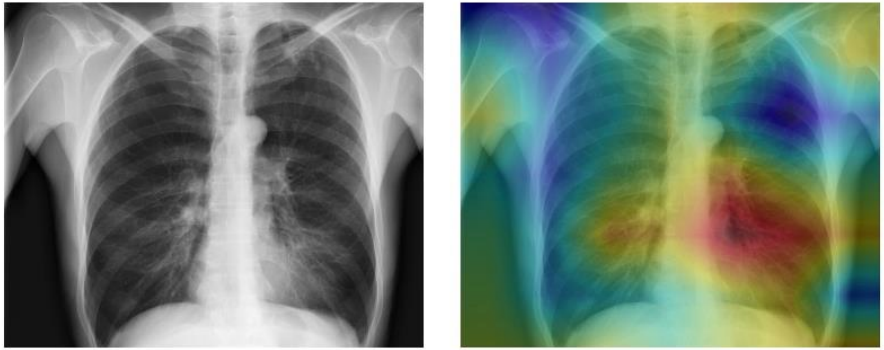
Videos Srinivasan et al. 2017

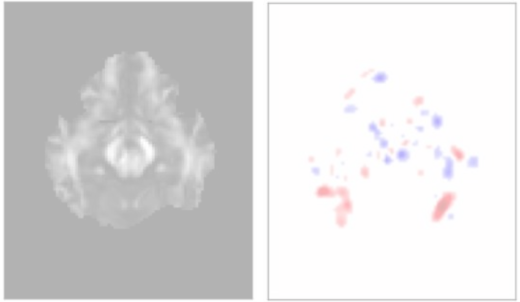
Audio Becker et al. 2019

. . .

Chest X-ray
Rajpurkar et al. 2017

Output
Pneumonia Positive (85%)





MRI brain scans Zintgraf et al. 2017

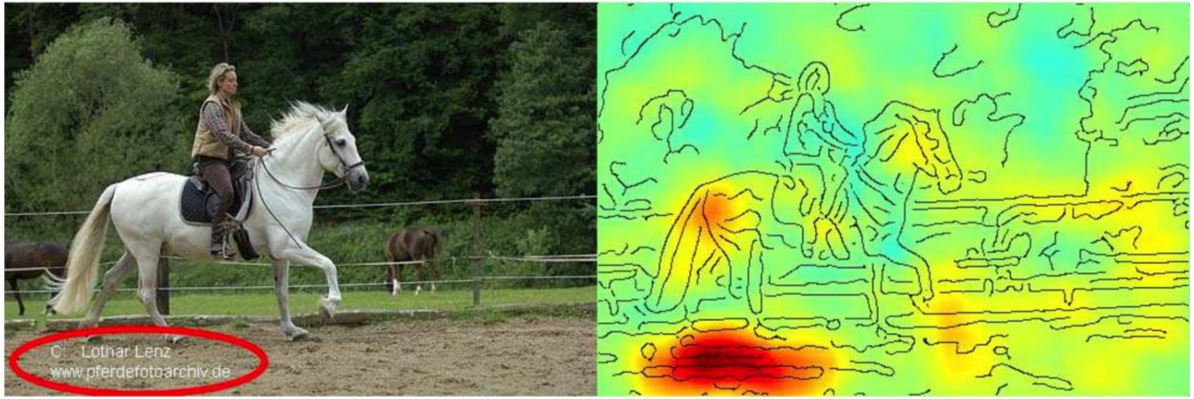
Video From 'wave'

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darw
Organization: University of New
Lines: 11
NNTP-Posting-Host: triton.unm.e

Text Ribeiro et al. 201

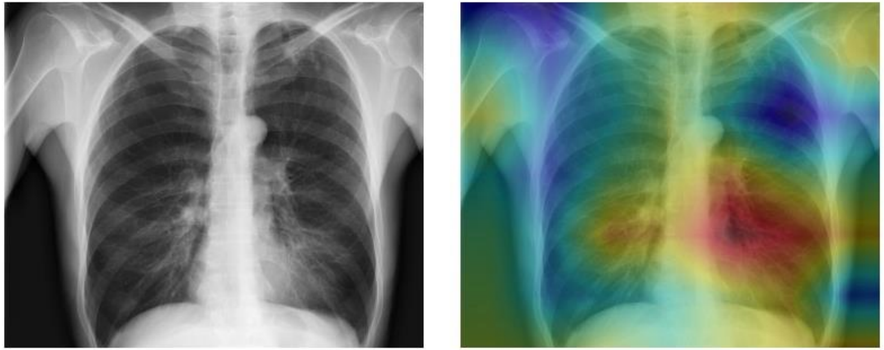
FIDO: Explaini
Expected Grac
FG-Vis: Interpretable and Fine-Grained Visual E
Understanding Deep Networks via Extremal Pe
MP-G: Removing input features via a generativ

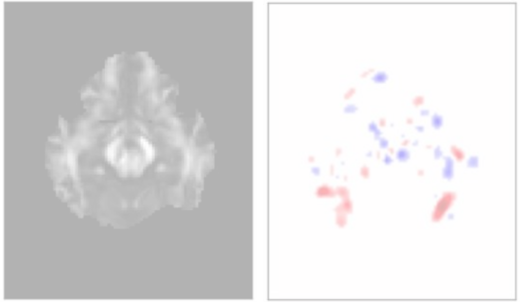
Detecting biases Lapuschkin et al. 2016



Chest X-ray Rajpurkar et al. 2017

Output Pneumonia Positive (85%)

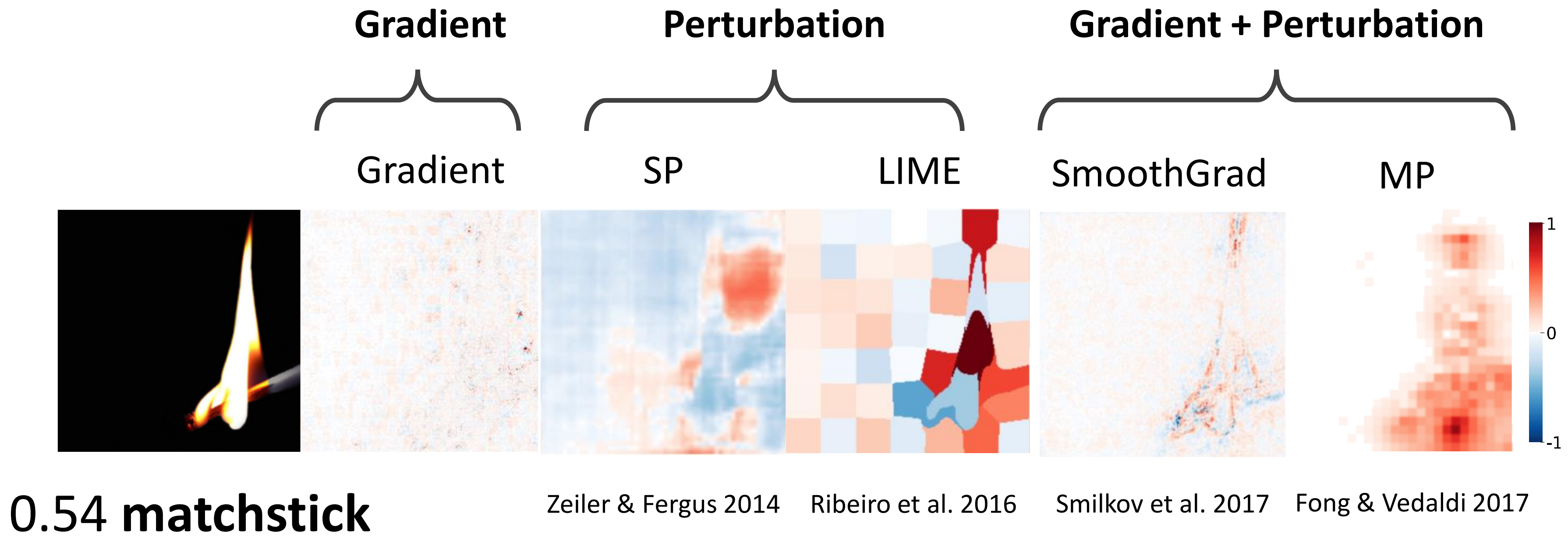




MRI brain scans Zintgraf et al. 2017

Detecting biases Lapuschkin et al. 2016

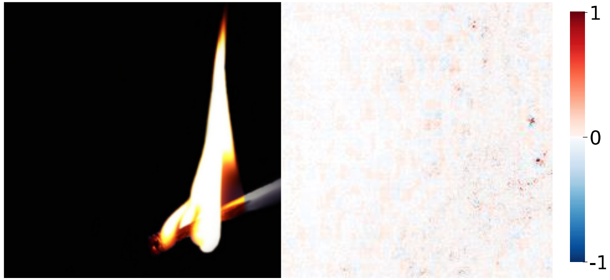
Are these explanations correct and reliable?



Are these explanations correct and reliable?

Method 0: Saliency maps

Gradient



$$\boxed{\nabla_{\boldsymbol{x}} f}$$

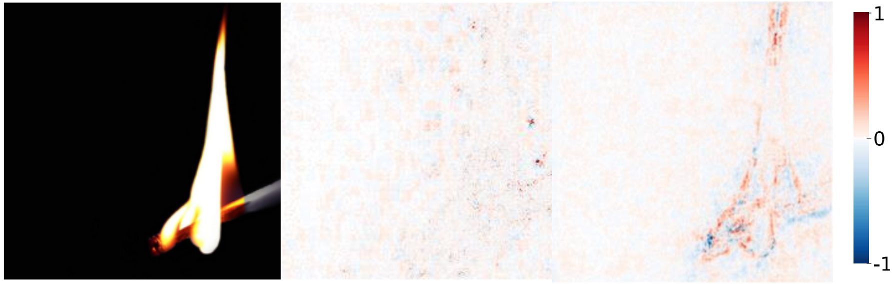
Problems:

- $\boxed{\text{too noisy}}$

Method 1: Smoothed saliency maps

Smilkov et al. 2017

Gradient SmoothGrad



$$\frac{1}{n}\sum_{1}^{n}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}+\mathcal{N}(0,\sigma^2))$$

Problems:

- ~~too noisy~~

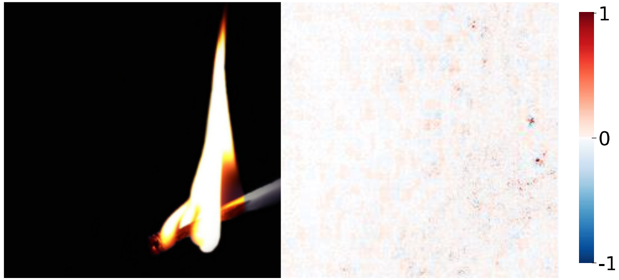
input model

$$A(\boldsymbol{x}, f, \mathcal{H})$$

$\mathcal{H} = \{n, \sigma\}$ hyperparameters

#1: Saliency maps may NOT be too noisy!

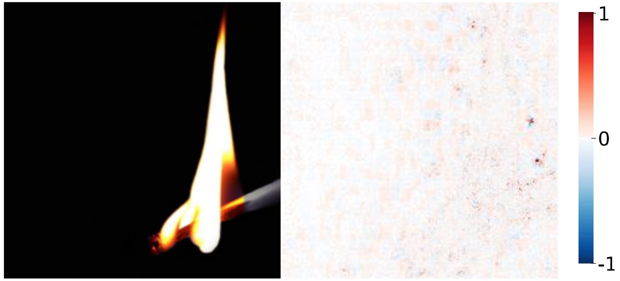
Gradient



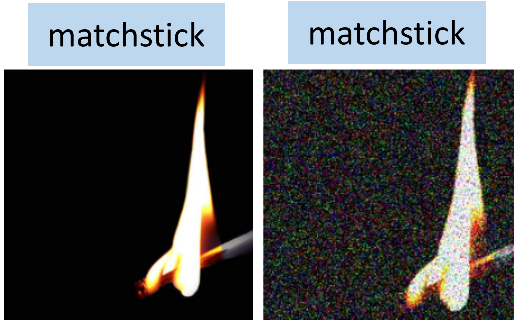
GoogLeNet

#1: Saliency maps may NOT be too noisy!

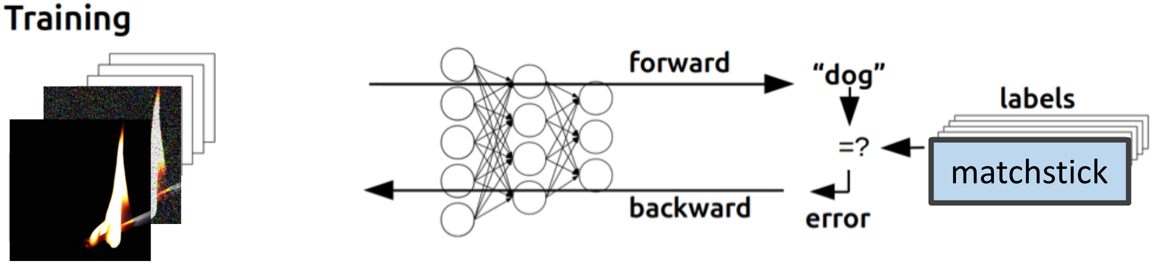
Gradient



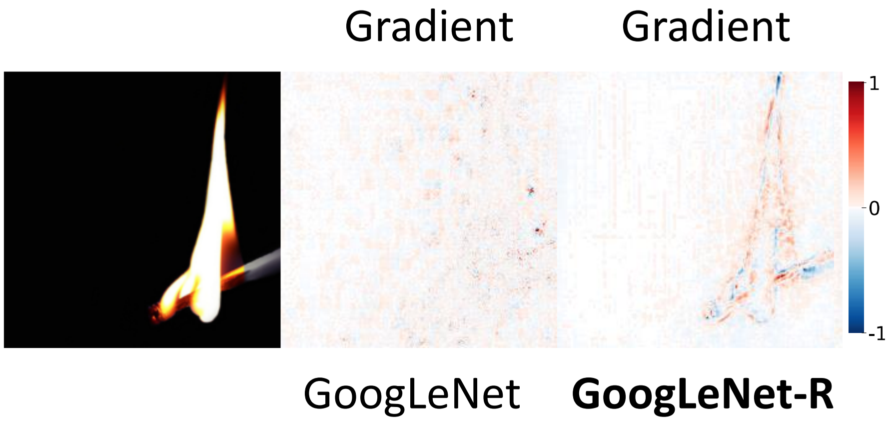
GoogLeNet GoogLeNet-R

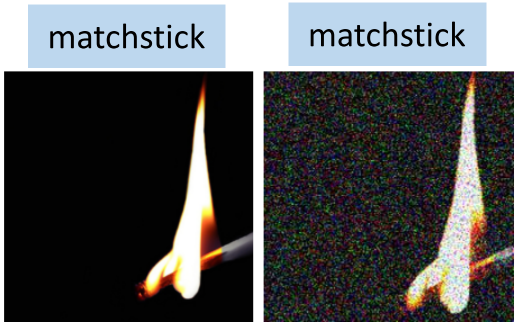


A *robust* classifier i.e.
adversarially trained with noisy images

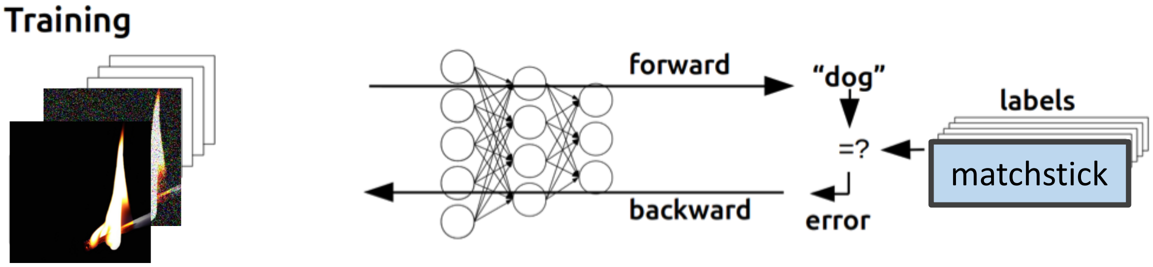


#1: Saliency maps may NOT be too noisy!



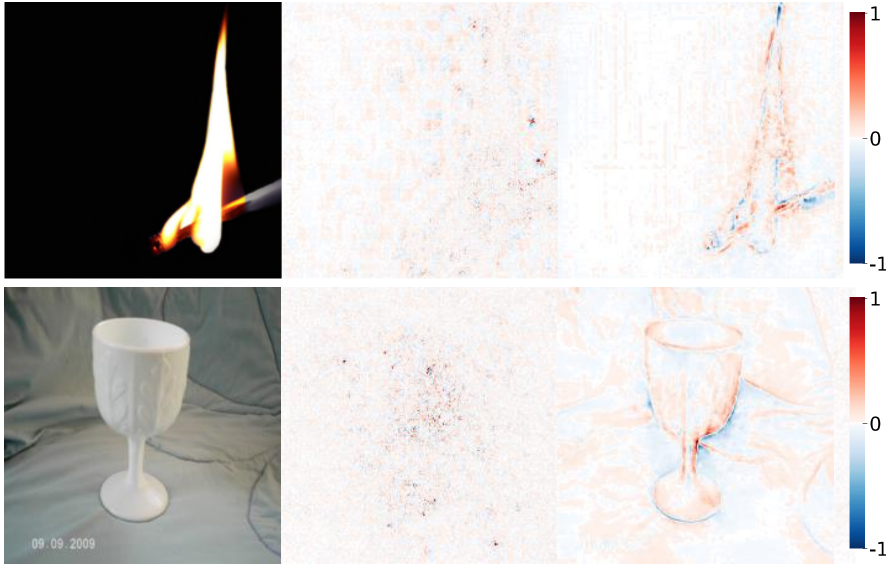


A *robust* classifier i.e.
adversarially trained with noisy images



#1: Saliency maps may NOT be too noisy!

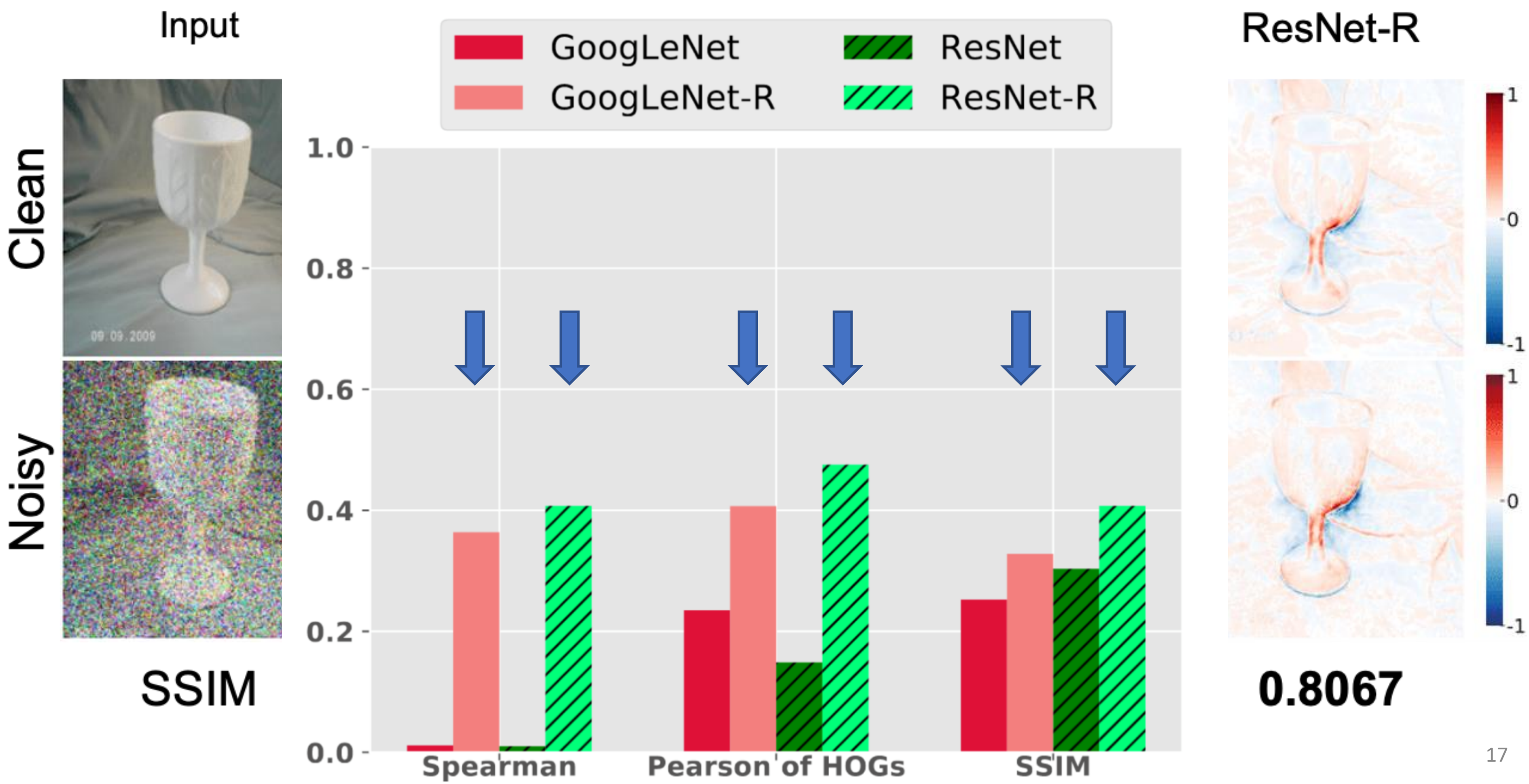
Gradient Gradient



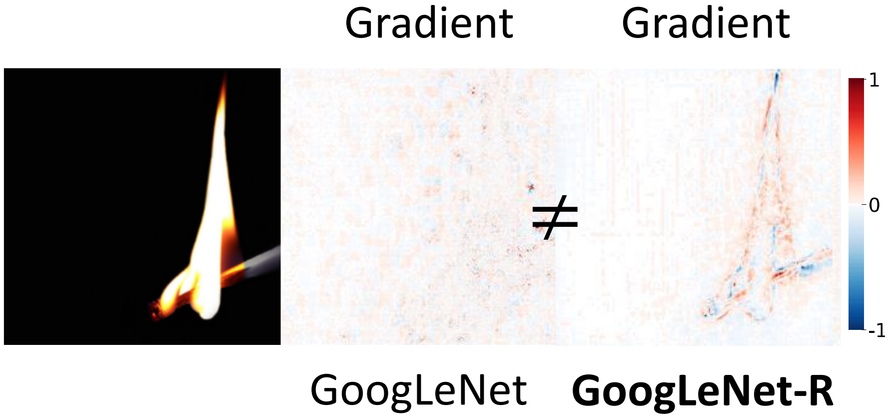
GoogLeNet **GoogLeNet-R**

A *robust* classifier i.e.
adversarially trained with noisy images

#1.1 Robust models are able to handle the additive noise to the input image



#1: Saliency maps may NOT be too noisy!



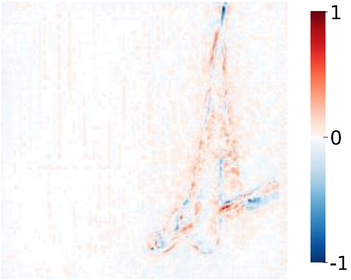
A *robust* classifier i.e.
adversarially trained with noisy images

#2: Smoothed gradients can be misinterpreted

Gradient

GoogLeNet

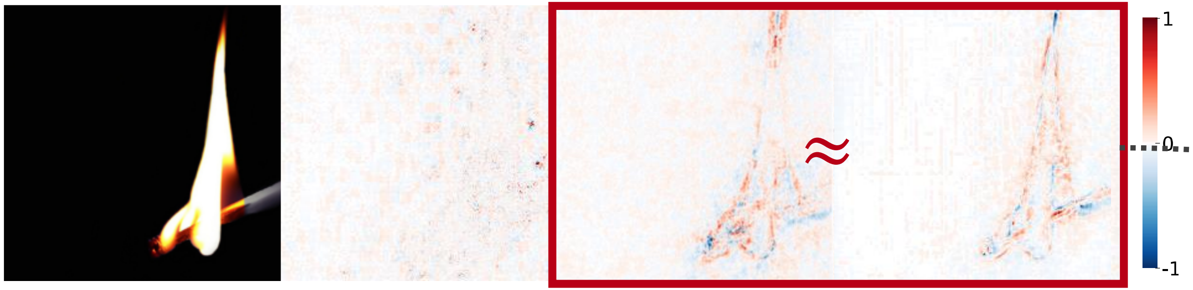
Gradient



GoogLeNet-R

A *robust* classifier i.e.
adversarially trained with noisy images

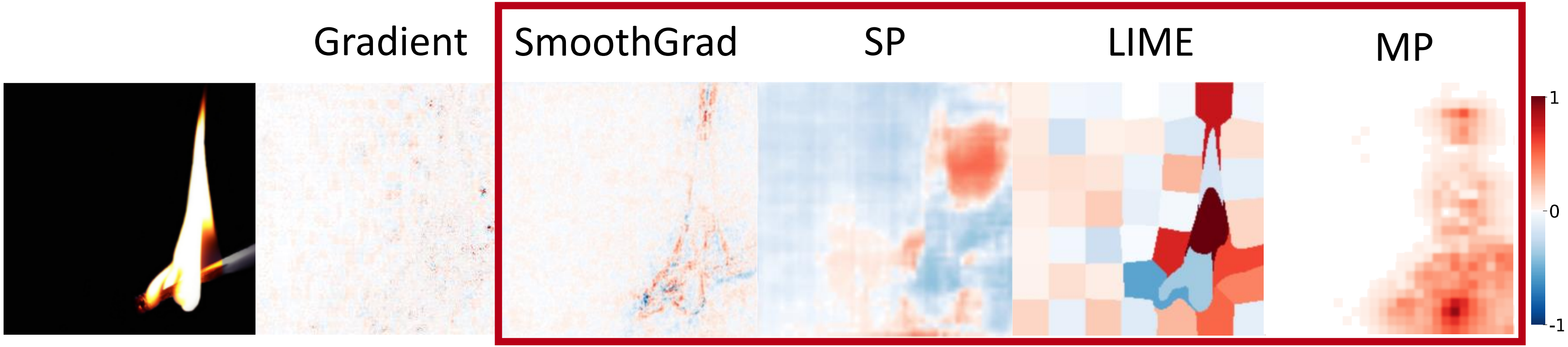
#2: Smoothed gradients can be misinterpreted



The **flame** is the most “important”

Some background pixels are the most “important”

#3: Many attribution maps are sensitive to hyperparams



input

model

$$A(\boldsymbol{x}, f, \boxed{\mathcal{H}})$$

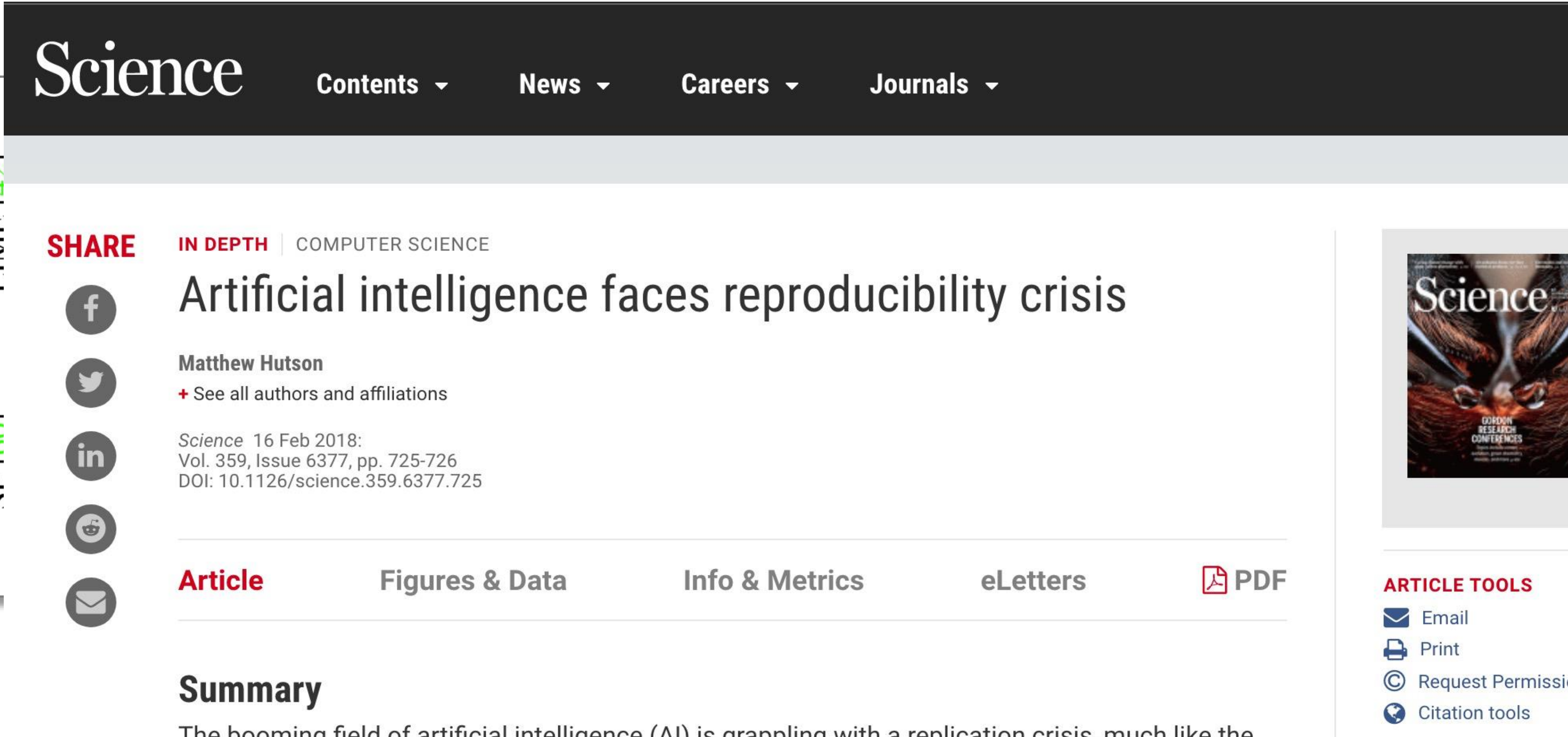
hyperparameters

#3: Many attribution maps are sensitive to hyperparams

Gradient SmoothGrad SP LIME MP

hyperparameters

#3: Many attribution maps are sensitive to hyperparams



Science Contents News Careers Journals

SHARE

IN DEPTH | COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Matthew Hutson

+ See all authors and affiliations

Science 16 Feb 2018:
Vol. 359, Issue 6377, pp. 725-726
DOI: 10.1126/science.359.6377.725

Article | Figures & Data | Info & Metrics | eLetters | PDF

ARTICLE TOOLS

Email

Print

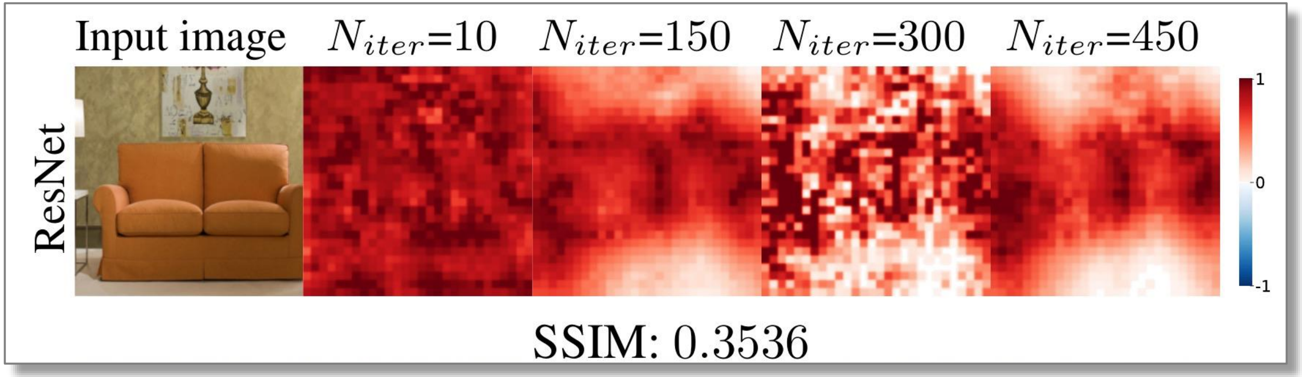
Request Permissi

Citation tools

Summary

The booming field of artificial intelligence (AI) is grappling with a replication crisis, much like the

#4: Attribution maps are more robust under *robust* classifiers

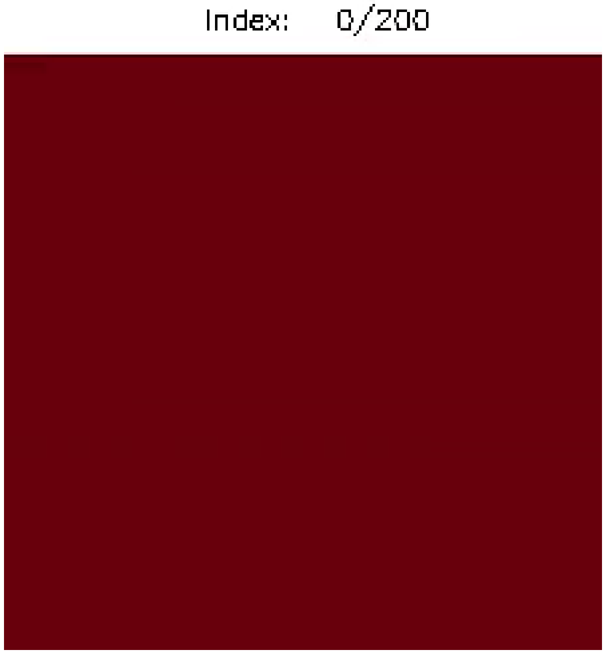


Idea: Find a minimal region s.t. when blurred out would minimize classification score

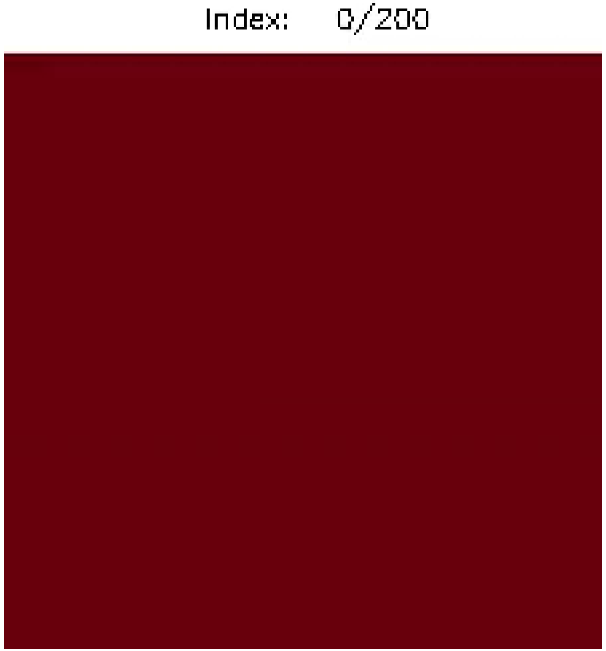
$$\boldsymbol{m}^* = \arg\min_{\boldsymbol{m}} \lambda \left\|\boldsymbol{m}\right\|_1 + f(\mathrm{blur}(\boldsymbol{x}, \boldsymbol{m}))$$

Meaningful-Perturbation (MP)

#4: Attribution maps are more robust under *robust* classifiers



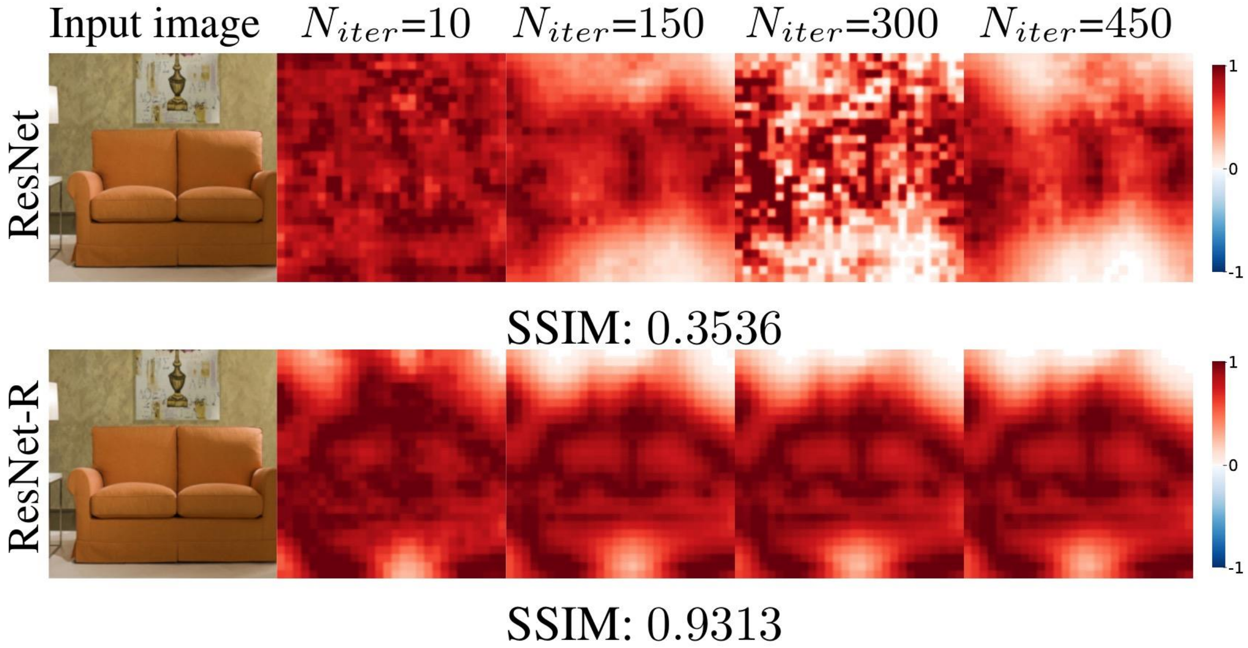
ResNet



ResNet-R

Meaningful-Perturbation (MP)

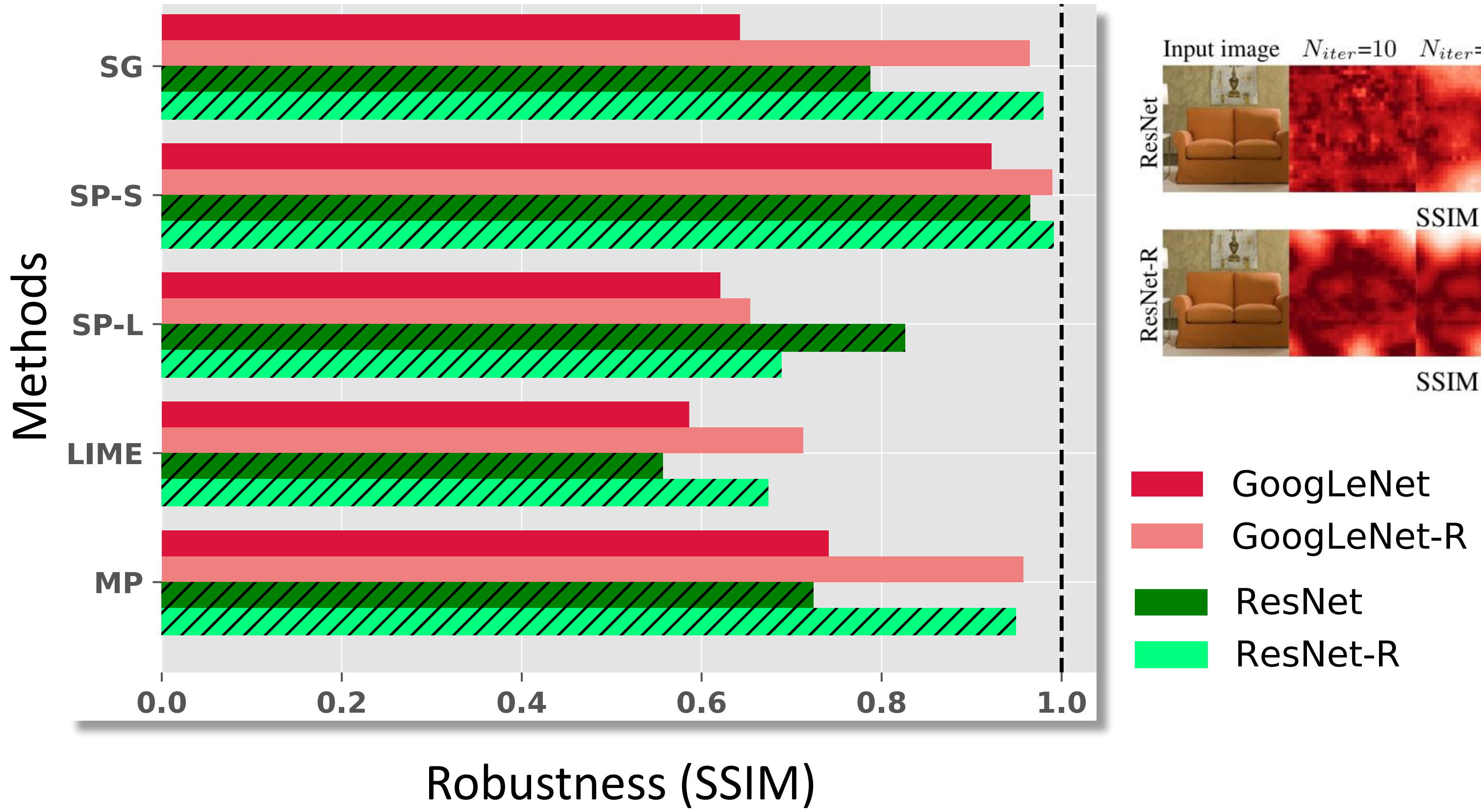
#4: Attribution maps are more robust under *robust* classifiers

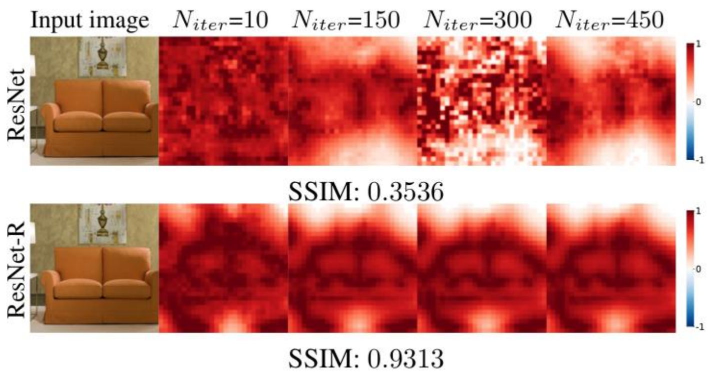


(b) Sensitivity to changes in the number of iterations N_{iter}

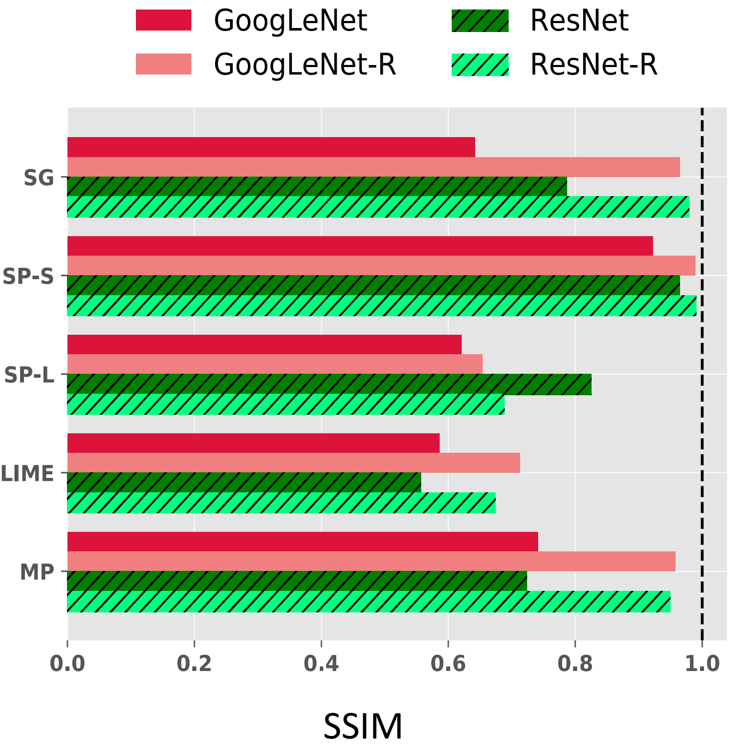
Meaningful-Perturbation (MP)

#4: Attribution maps are more robust under *robust* classifiers



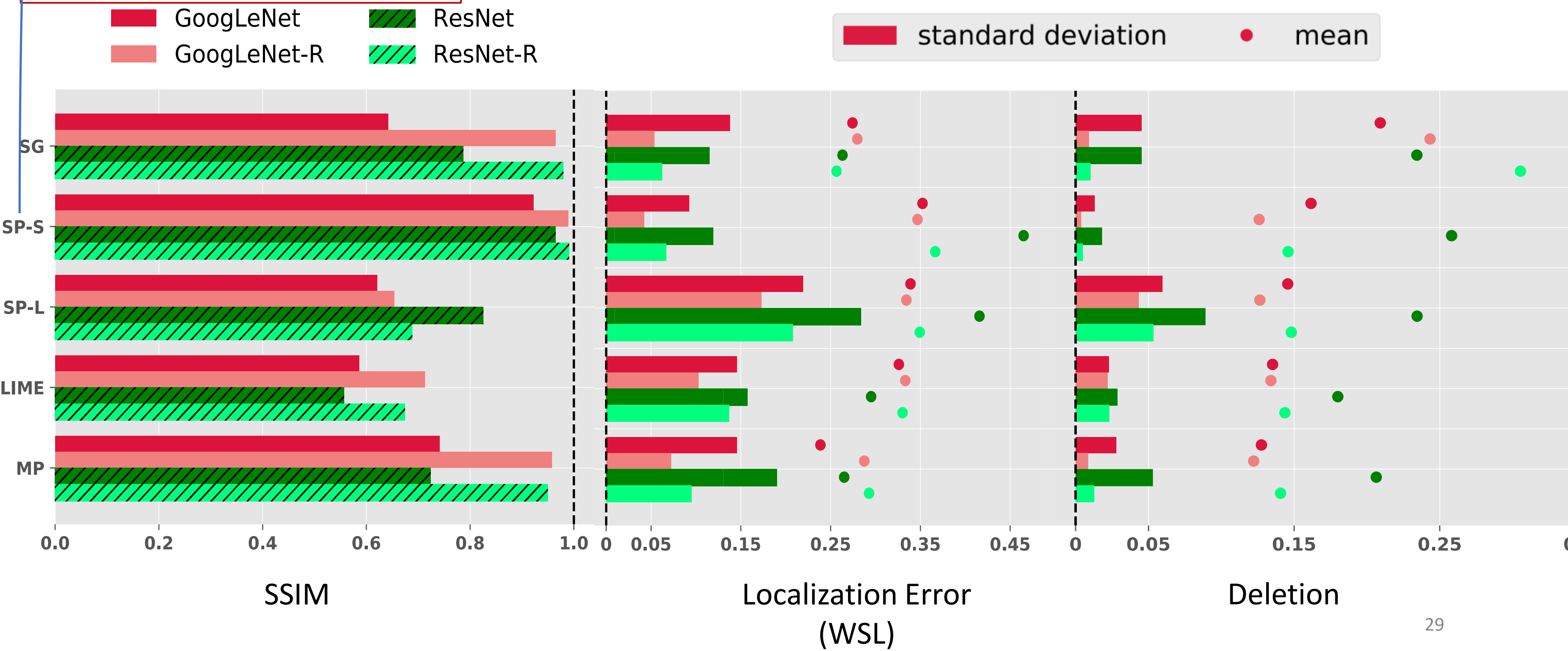


#5 Pixel-wise sensitivity translates to sensitivity in accuracy scores/downstream tasks

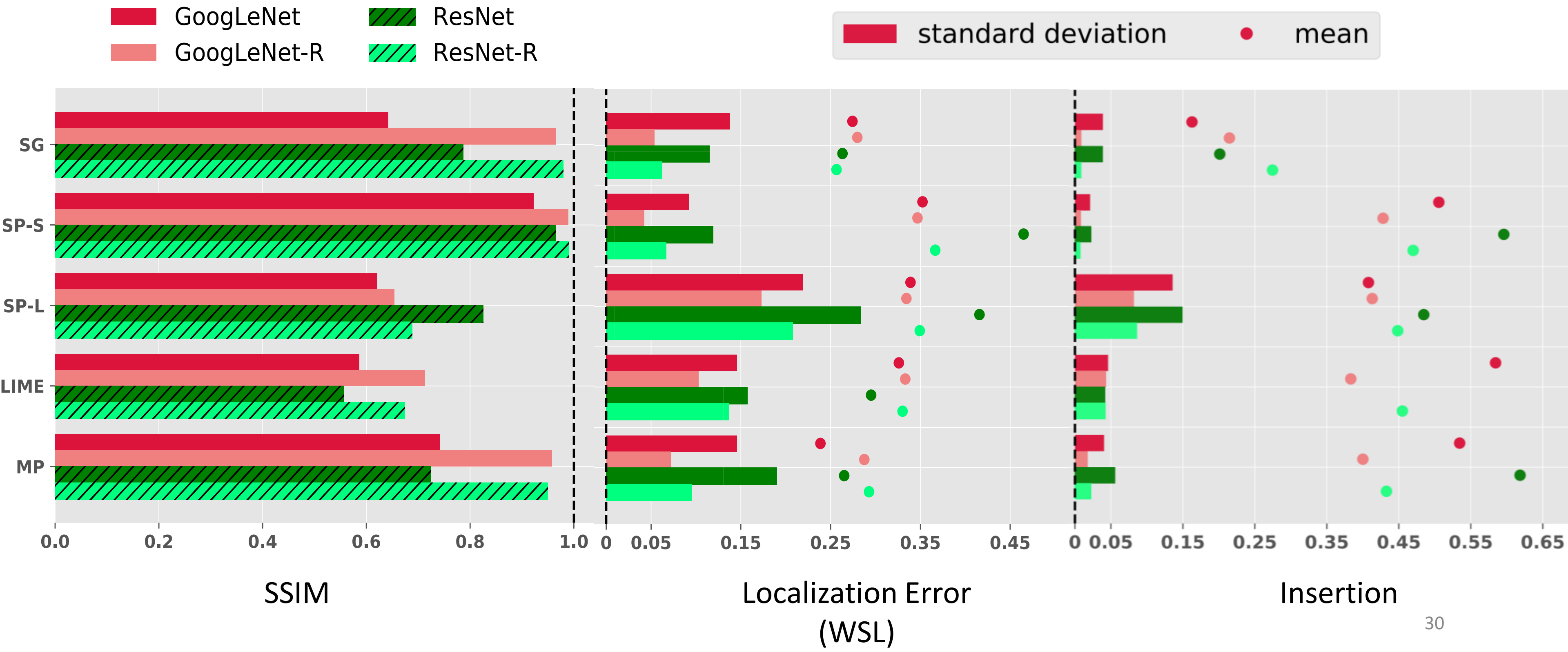


#5 Pixel-wise sensitivity translates to sensitivity in accuracy scores/downstream tasks

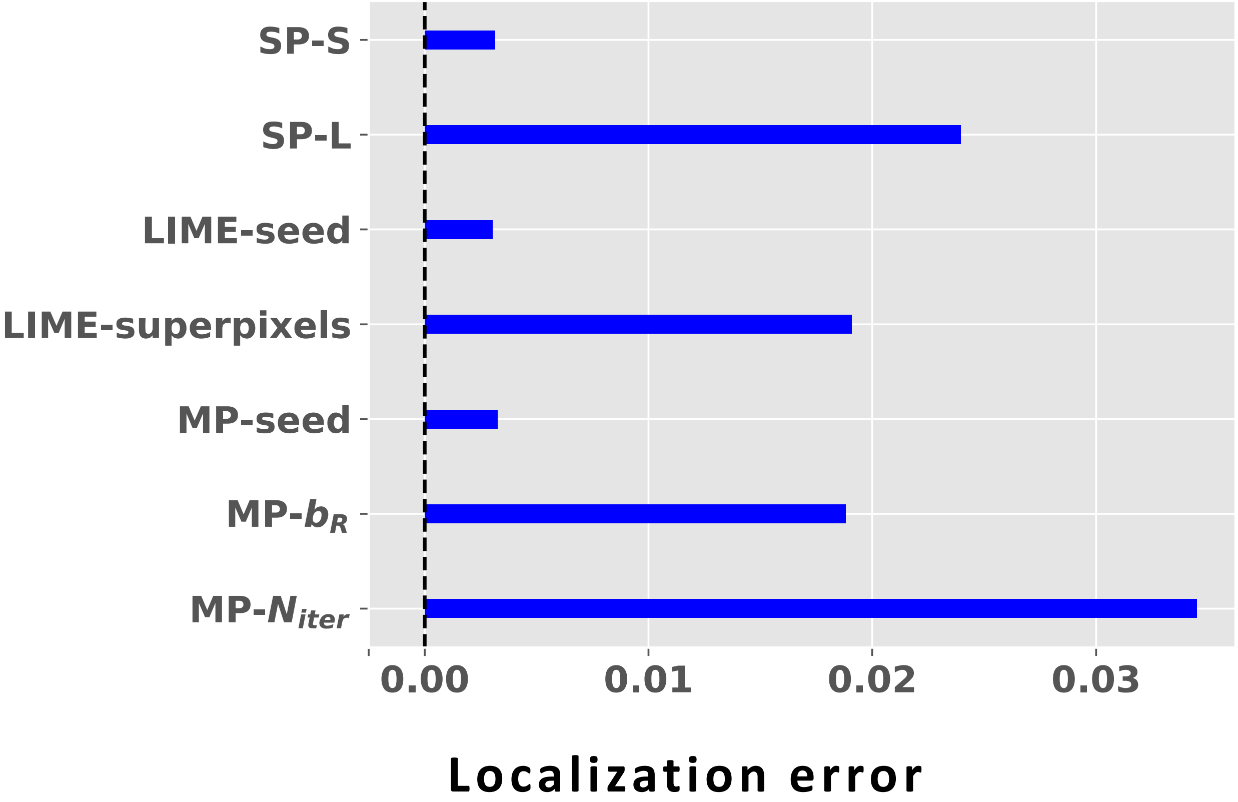
Sliding Patch with very close patch sizes (52, 53, 54)



#5 Pixel-wise sensitivity translates to sensitivity in accuracy scores/downstream tasks



#6: Some hyperparameters are more detrimental



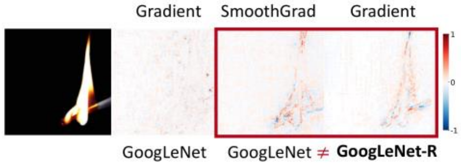
Conclusions


Naman

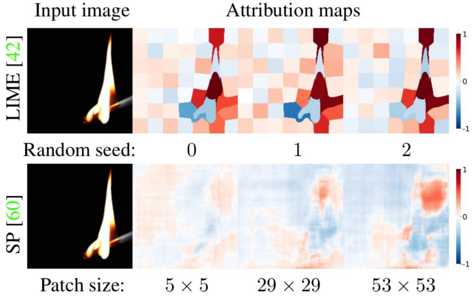

Chirag

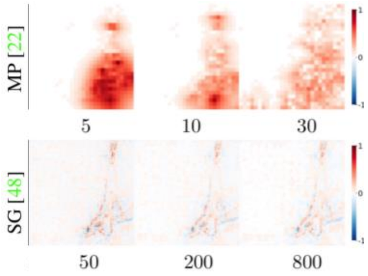

Anh

1. Gradient images for robust classifiers are smooth
2. Smoothing gradients may cause misinterpretation
3. Many attribution methods are sensitive to hyper-parameters
4. For robust classifiers, attribution maps are more robust



Paper & code:
http://anhnguyen.me/project/sam





Work funded by