How to explain neural network decisions?

Anh Nguyen

Assistant Professor

@anh_ng8
http://AnhNguyen.me

Challenge: Out-of-distribution Generalization

– Leon Bottou (ICML 2019 Keynote)
– Yoshua Bengio (NeurIPS 2019 Keynote, 2019 AI Debate)

ImageNet
school bus class

Task: Image classification

Inception-v3
78% accuracy on ImageNet

100%
school bus

0% school bus!

Alcorn et al. CVPR 2019

Inception-v3
78% accuracy on ImageNet

100%
school bus

100% confidence

Alcorn et al. CVPR 2019

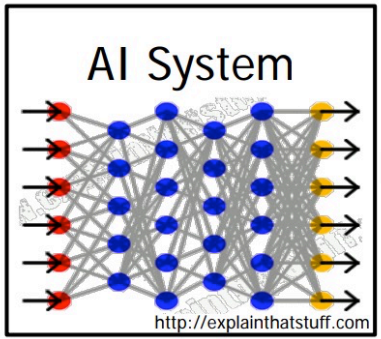
vacuum cleaner

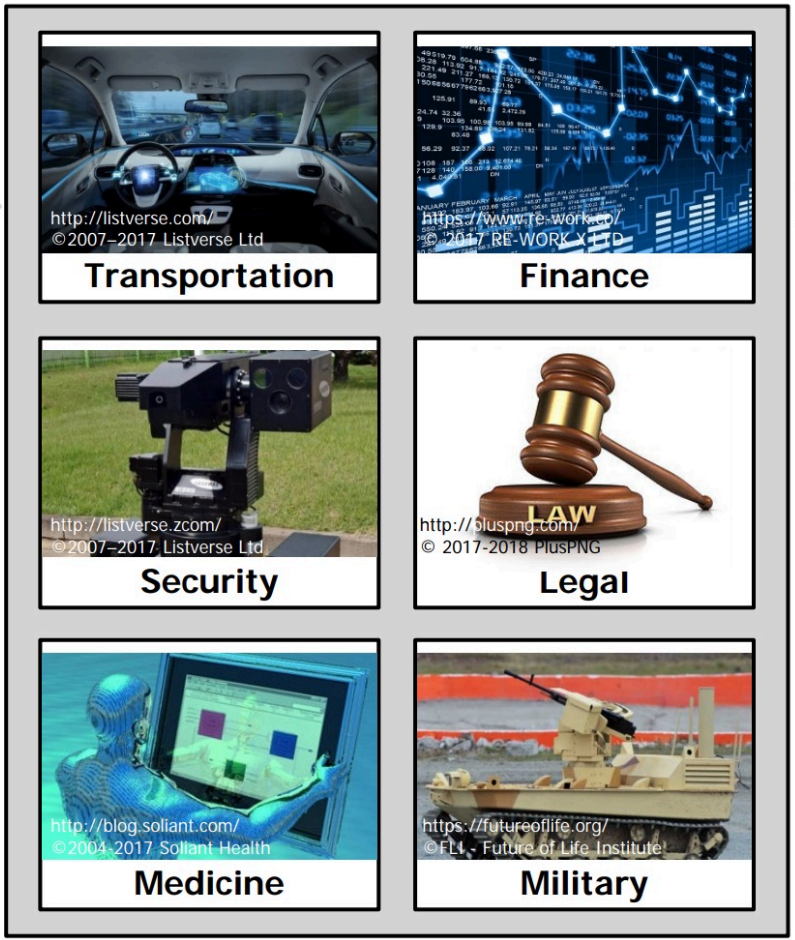
Challenge: Out-of-distribution Generalization

shopping basket

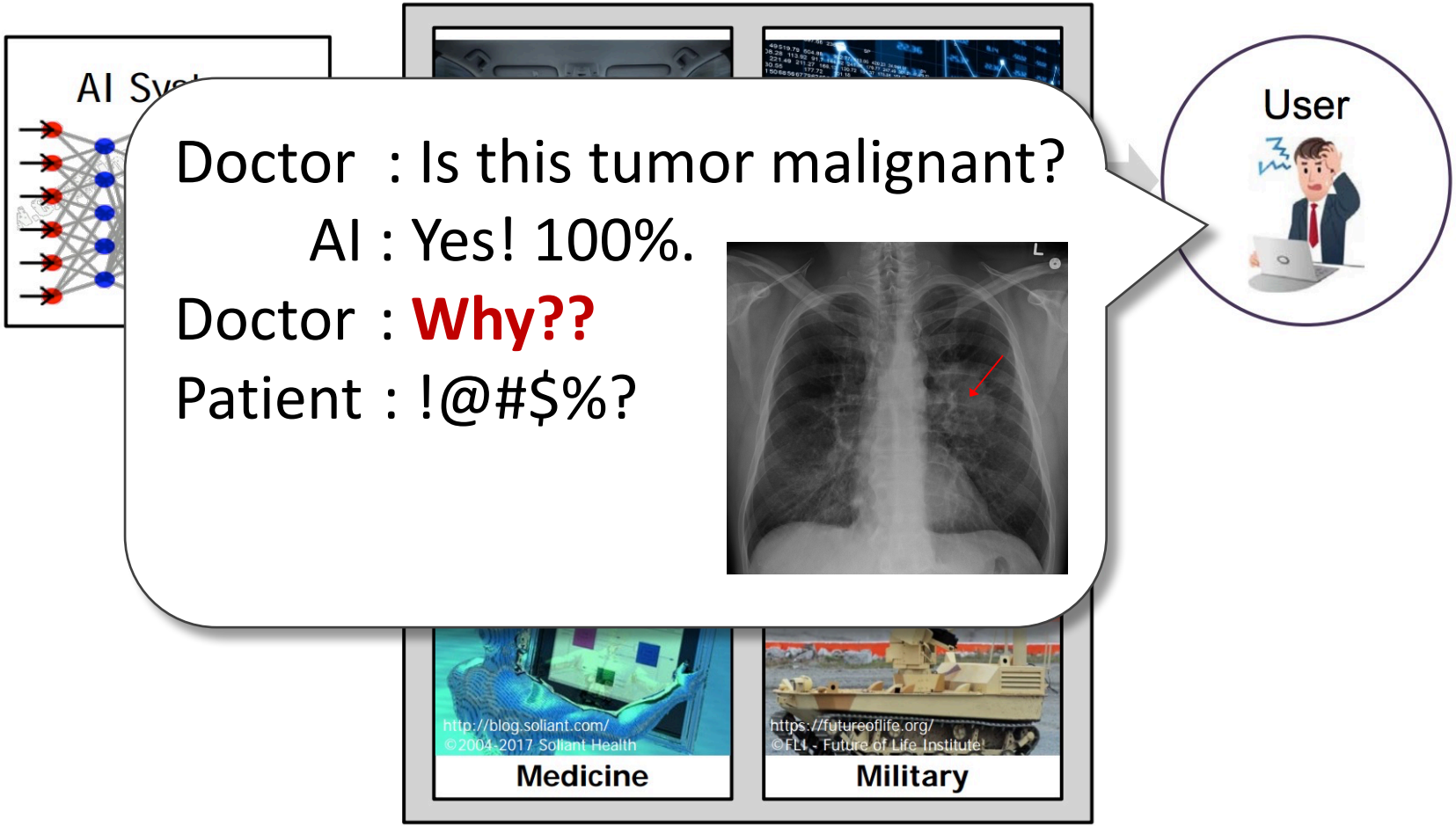
Alcorn et al. CVPR 2019

Paper & code: http://anhnguyen.me/project/strike-with-a-pose/

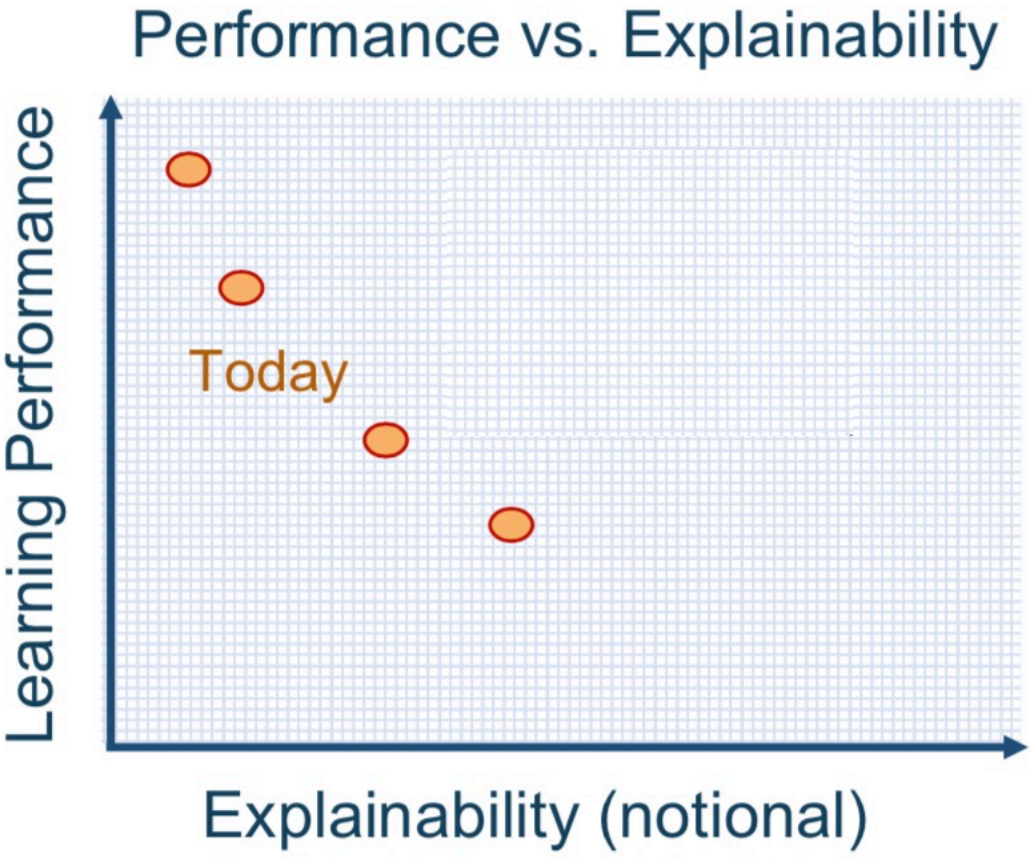




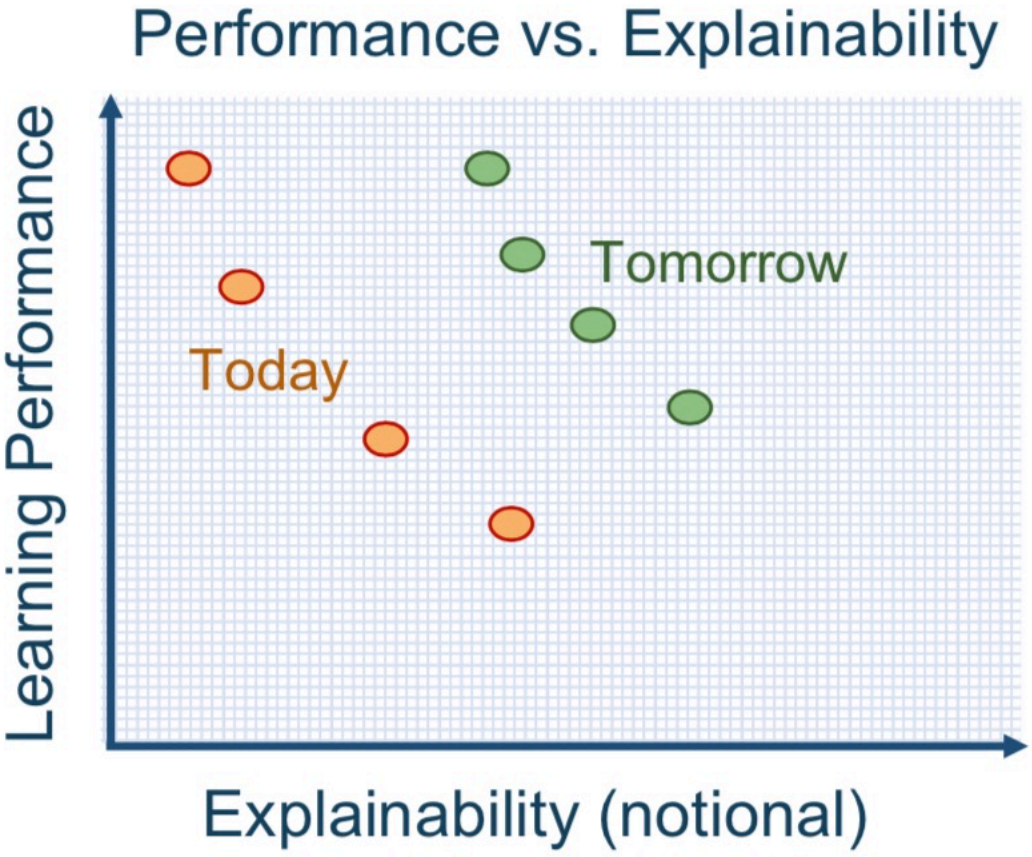
https://www.darpa.mil/attachments/XAIProgramUpdate.pdf



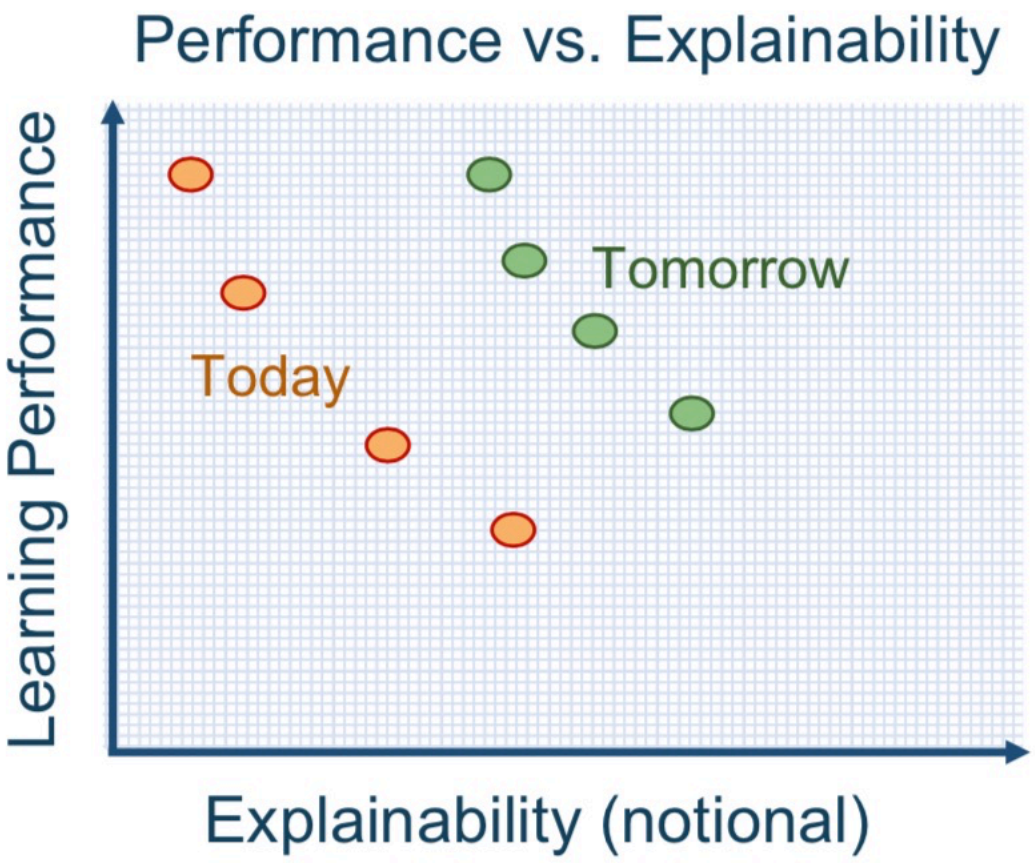
https://www.darpa.mil/attachments/XAIProgramUpdate.pdf



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



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



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

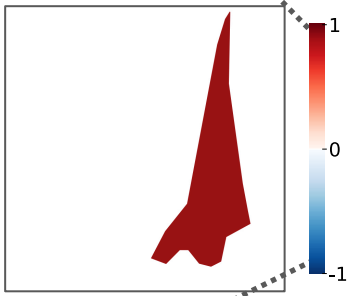
Measure of Explanation Effectiveness
User Satisfaction
• Clarity of the explanation (user rating) • Utility of the explanation (user rating)
Mental Model
• Understanding individual decisions • Understanding the overall model • Strength/weakness assessment • 'What will it do' prediction • 'How do I intervene' prediction
Task Performance
• Does the explanation improve the user's decision, task performance? • Artificial decision tasks introduced to diagnose the user's understanding
Trust Assessment
• Appropriate future use and trust
Correctability (Extra Credit)
• Identifying errors • Correcting errors • Continuous training

Attribution maps as explanations

x → f → $f(x)$

0.54 **matchstick**

for matchstick



against matchstick

attribution map
(hypothetical)

input

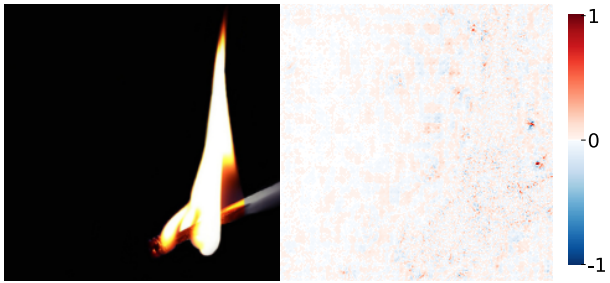
model

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

hyperparameters

Method 0: Saliency maps

Gradient



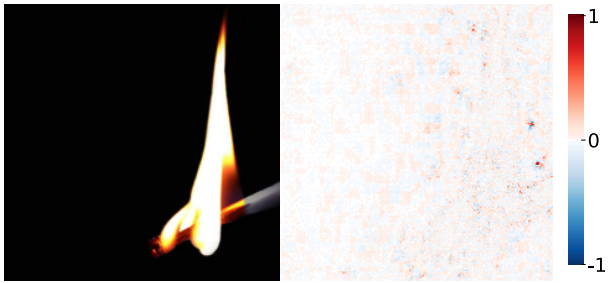
$$\nabla_{\boldsymbol{x}} f$$

Problems:

- too noisy

Method 1: Smoothed saliency maps

Gradient



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

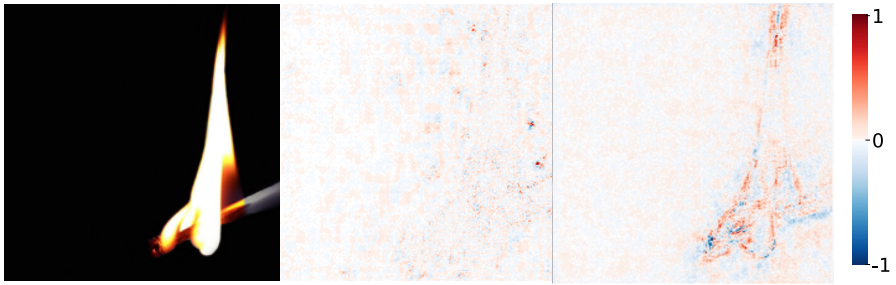
Problems:

- 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

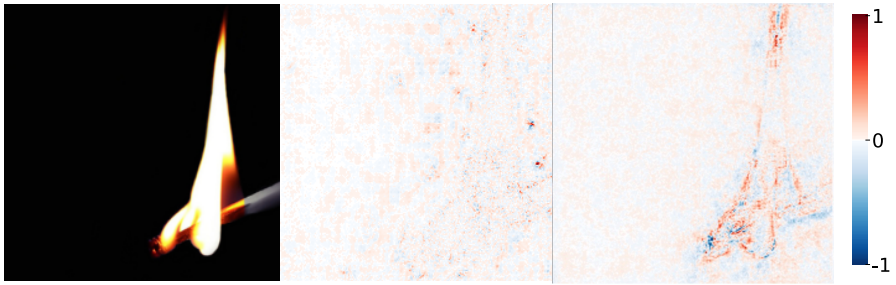
$$A(\boldsymbol{x}, f, \mathcal{H})$$

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

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~~
- sensitivity ≠ importance

input model

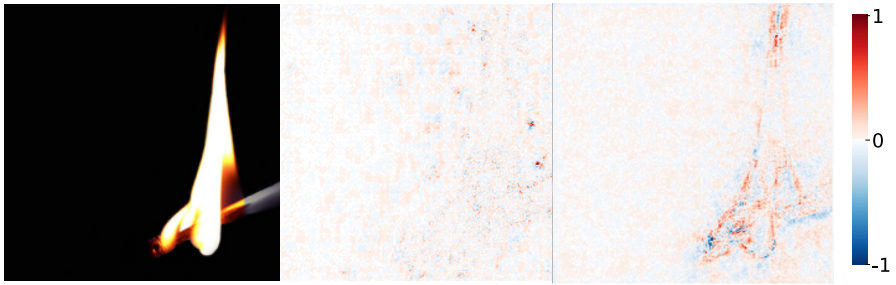
$$A(\boldsymbol{x}, f, \mathcal{H})$$

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

Method 1: Smoothed saliency maps

Smilkov et al. 2017

Gradient SmoothGrad



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

How does prediction change if the flame is slightly *whiter*?

How does prediction change if the flame is *NOT present*?

Problems:

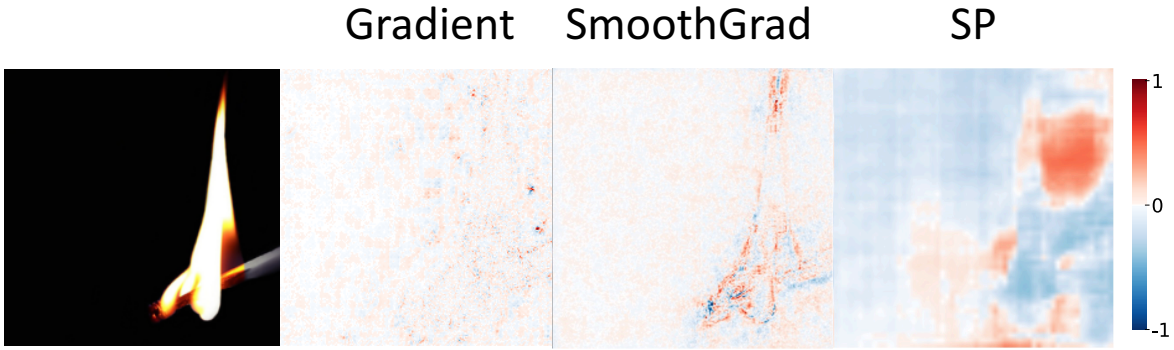
- ~~too noisy~~
- sensitivity ≠ importance

input model

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

Method 2: Sliding Patch

Zeiler & Fergus 2014



How does prediction change if the flame is *NOT present*?

importance

Method 2: Sliding Patch

Zeiler & Fergus 2014

How does prediction change if the flame is *NOT present*?

importance

Method 2: Sliding Patch

Zeiler & Fergus 2014

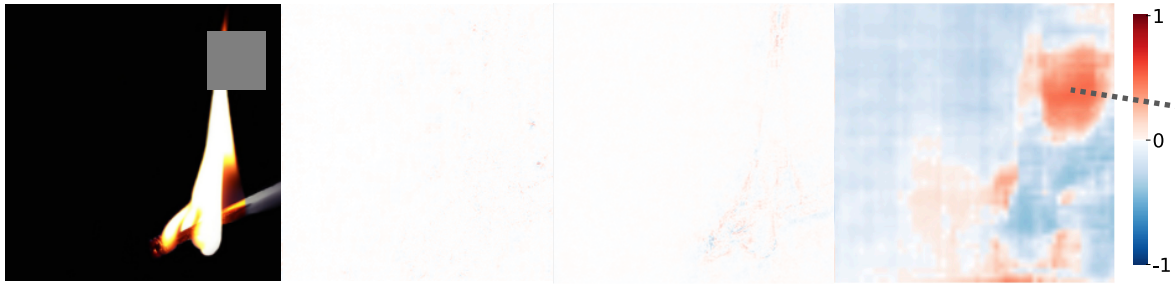
How does prediction change if the flame is *NOT present*?

importance

Method 2: Sliding Patch

Zeiler & Fergus 2014

$\bar{x}$ Gradient SmoothGrad SP



"treatment effect"

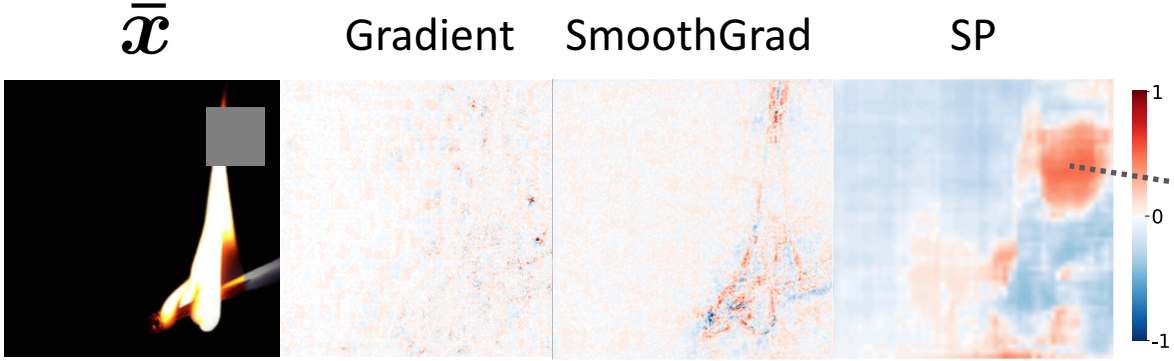
$$f(x) - f(\bar{x})$$

How does prediction change if the flame is *NOT present*?

importance

Method 2: Sliding Patch

Zeiler & Fergus 2014



"treatment effect"

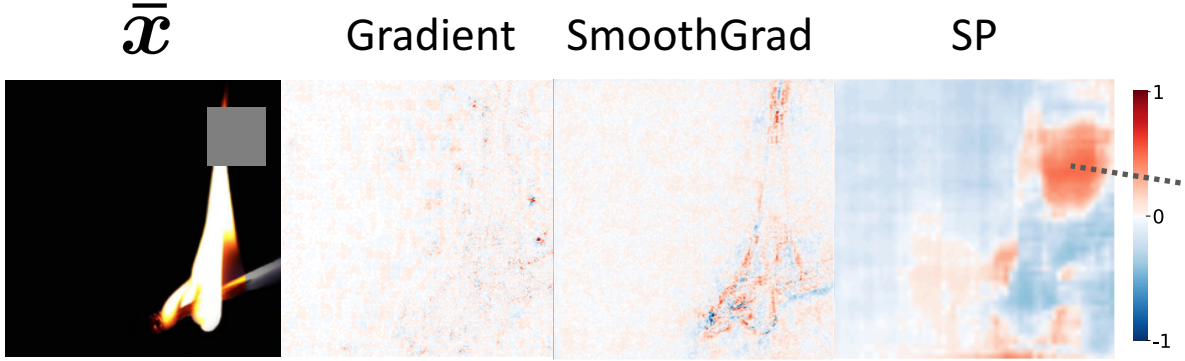
$$f(x) - f(\bar{x})$$

Problems:

- ~~sensitivity ≠ importance~~
- ~~too noisy~~

Method 2: Sliding Patch

Zeiler & Fergus 2014



"treatment effect"

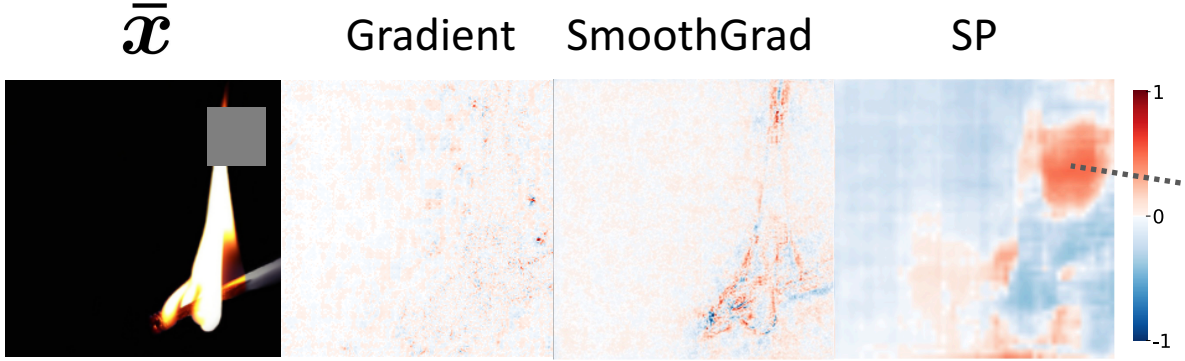
$$f(x) - f(\bar{x})$$

Problems:

- ~~too noisy~~
- ~~sensitivity ≠ importance~~
- coarse, inaccurate (due to square patch)

Method 2: Sliding Patch

Zeiler & Fergus 2014



"treatment effect"

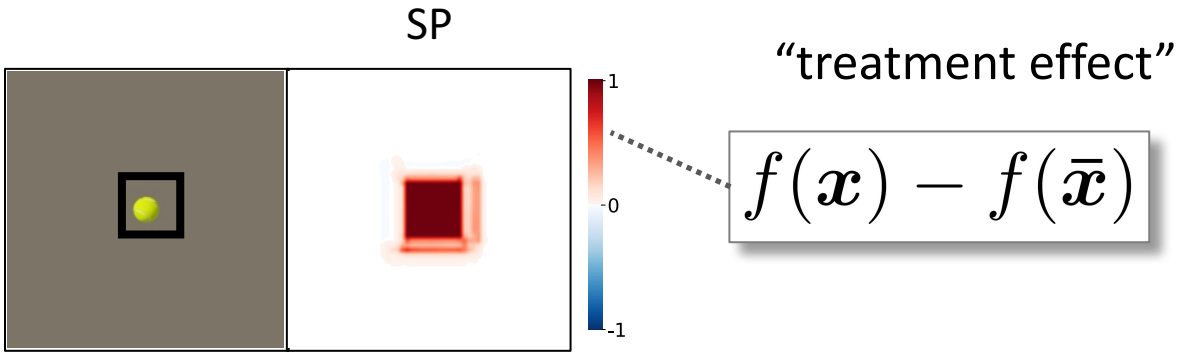
$$f(x) - f(\bar{x})$$

Problems:

- coarse, inaccurate (due to square patch)

Method 2: Sliding Patch

Zeiler & Fergus 2014

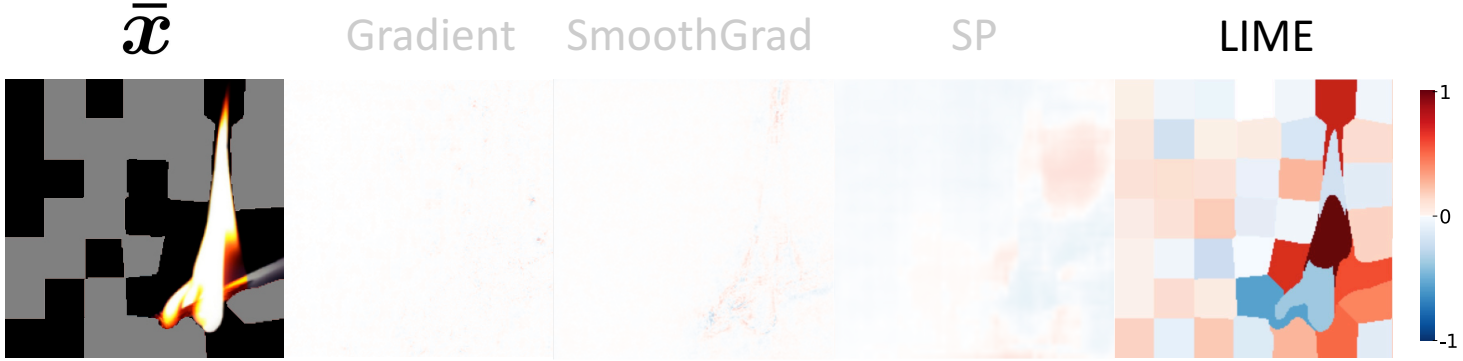


Problems:

- coarse, inaccurate (due to square patch)

Method 3: LIME

Ribeiro et al. 2016



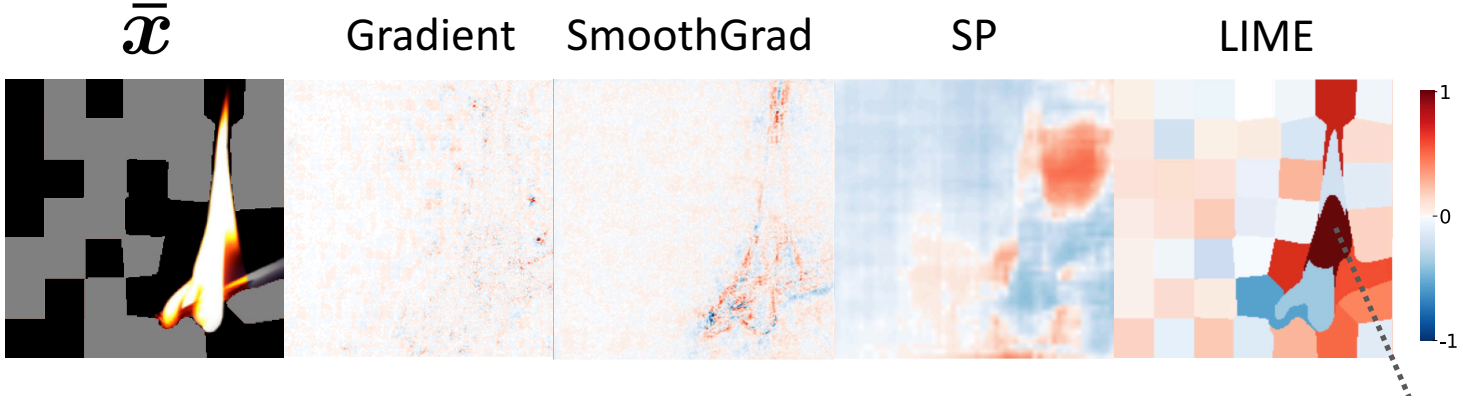
Idea: Attribution for superpixel k is the average score when k is visible

Problems:

- coarse, inaccurate (due to square patch)

Method 3: LIME

Ribeiro et al. 2016



Idea: Attribution for superpixel k is the average score when k is visible

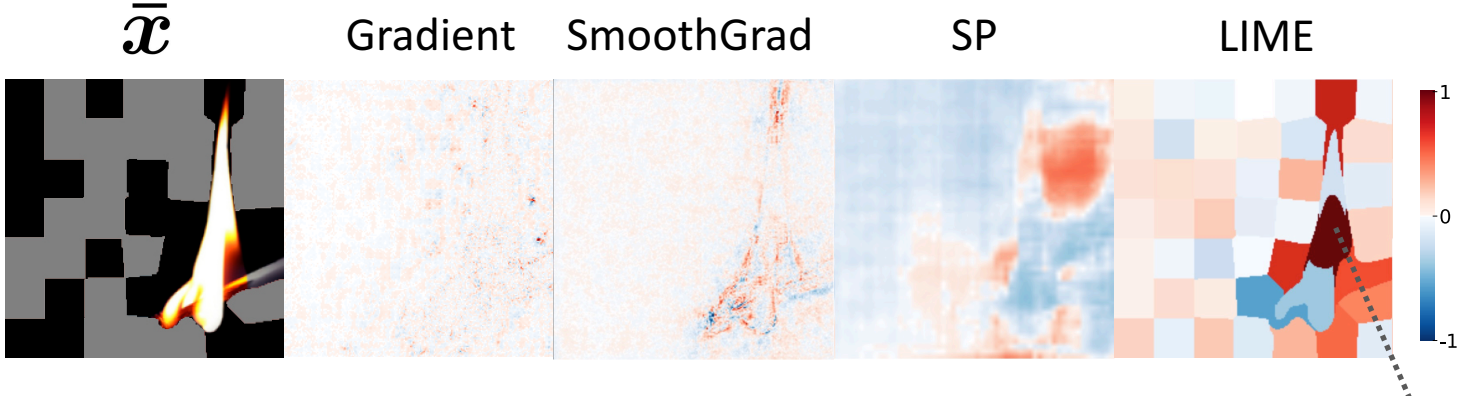
Problems:

- coarse ~~inaccurate (due to square patch~~)
- gray pixels?

$$\simeq \frac{1}{n}\sum_{i=1}^{n}\frac{f(\bar{\boldsymbol{x}}^i)}{S/2}$$

Method 3: LIME

Ribeiro et al. 2016



Idea: Attribution for superpixel k is the average score when k is visible

Problems:

- coarse
- gray pixels?

$$\simeq \frac{1}{n}\sum_{i=1}^{n}\frac{f(\bar{\boldsymbol{x}}^i)}{S/2}$$

Method 4: Learned blur mask

Fong & Vedaldi 2017

Problems:

- coarse
- gray pixels?

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

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

Method 4: Learned blur mask

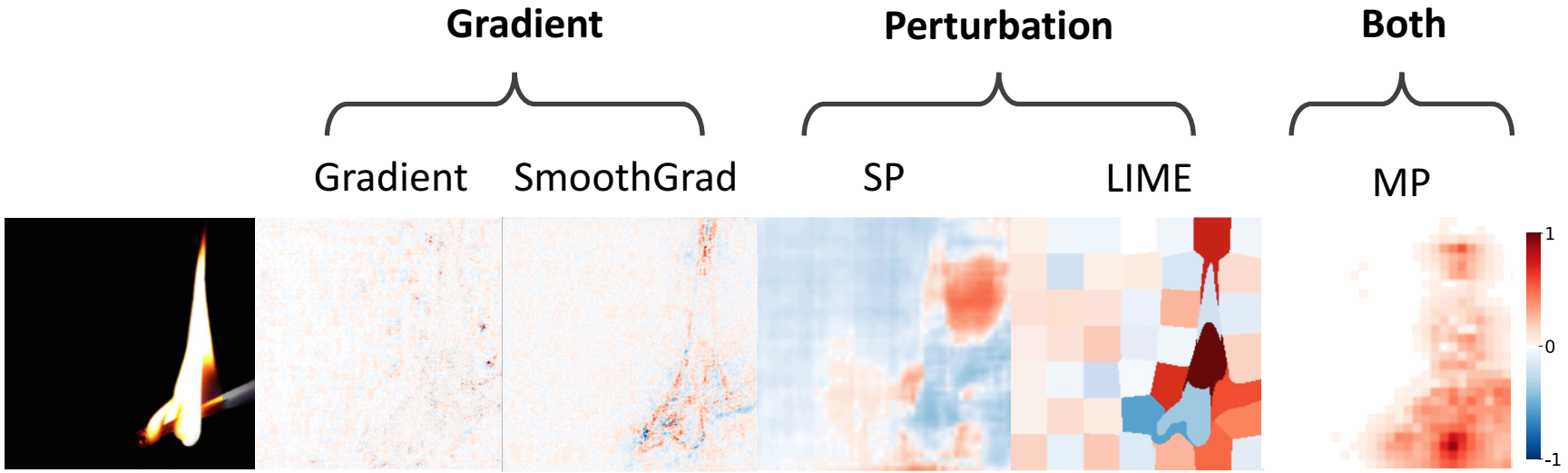
Fong & Vedaldi 2017

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

Problems:

- ~~coarse~~ → fine-grained
- ~~gray~~ → blurred pixels

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

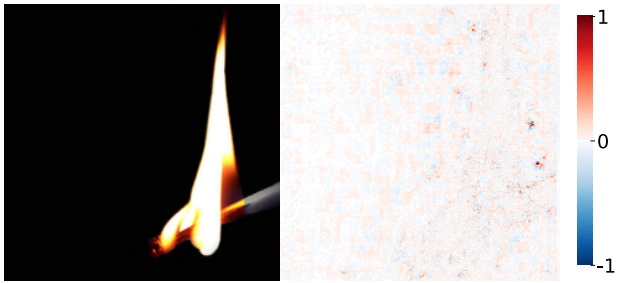


0.54 **matchstick**

Are these explanations correct and reliable?

#1: Saliency maps may NOT be too noisy!

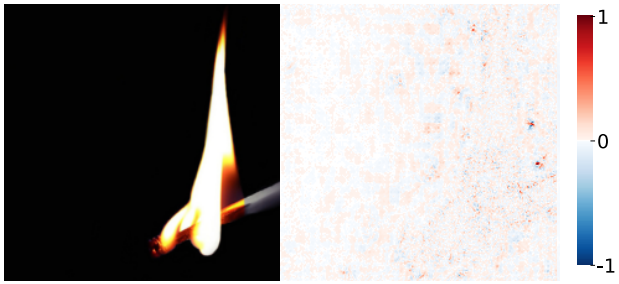
Gradient



GoogLeNet

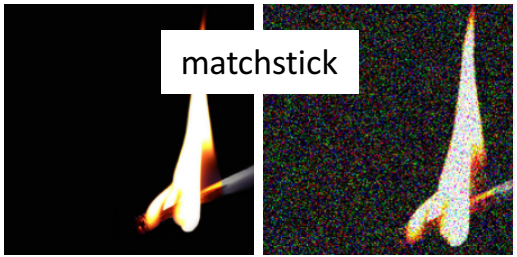
#1: Saliency maps may NOT be too noisy!

Gradient



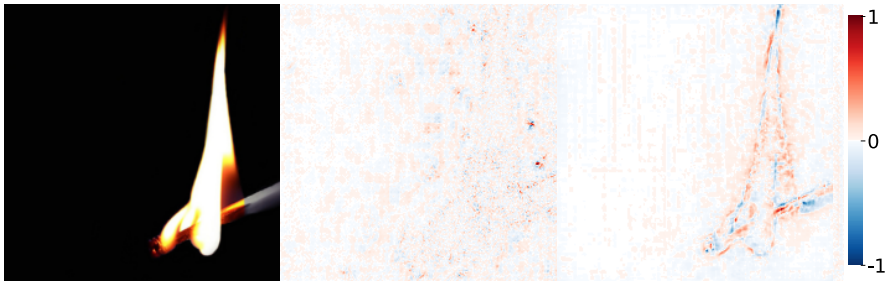
GoogLeNet **GoogLeNet-R**

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



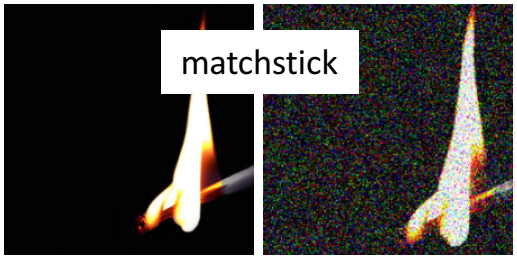
#1: Saliency maps may NOT be too noisy!

Gradient Gradient

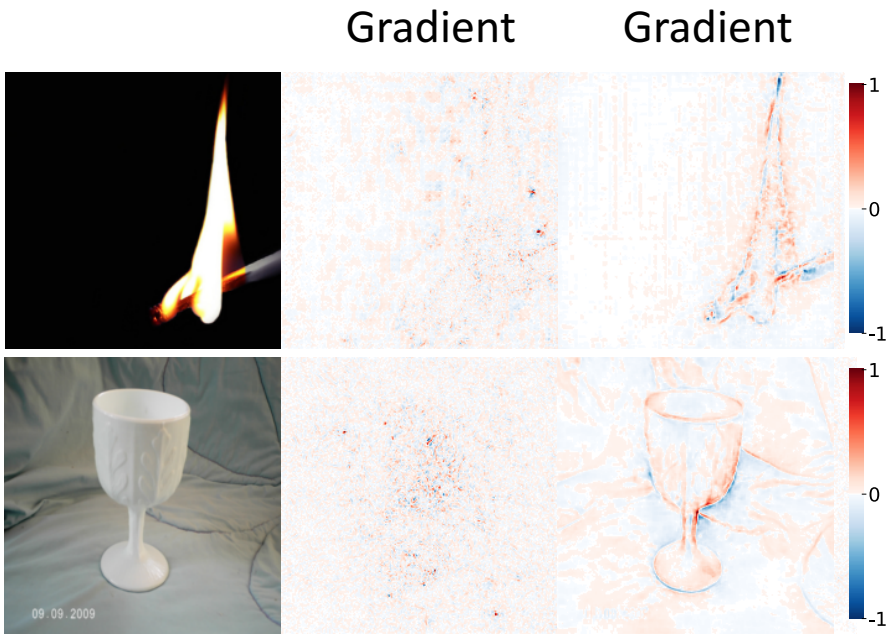


GoogLeNet **GoogLeNet-R**

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



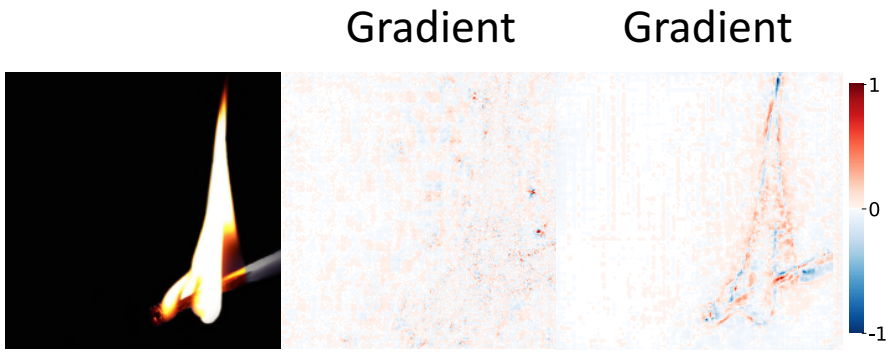
#1: Saliency maps may NOT be too noisy!



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

Bansal*, Agarwal*, Nguyen*, 2020

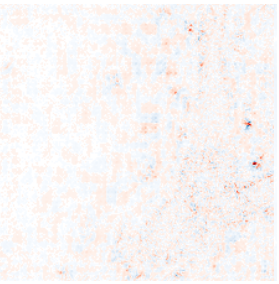
#1: Saliency maps may NOT be too noisy!



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

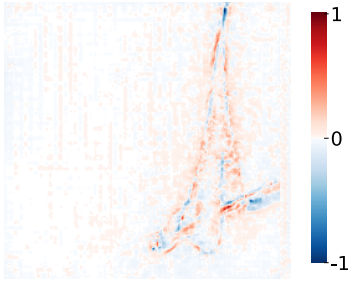
#1: Saliency maps may NOT be too noisy!

Gradient



GoogLeNet

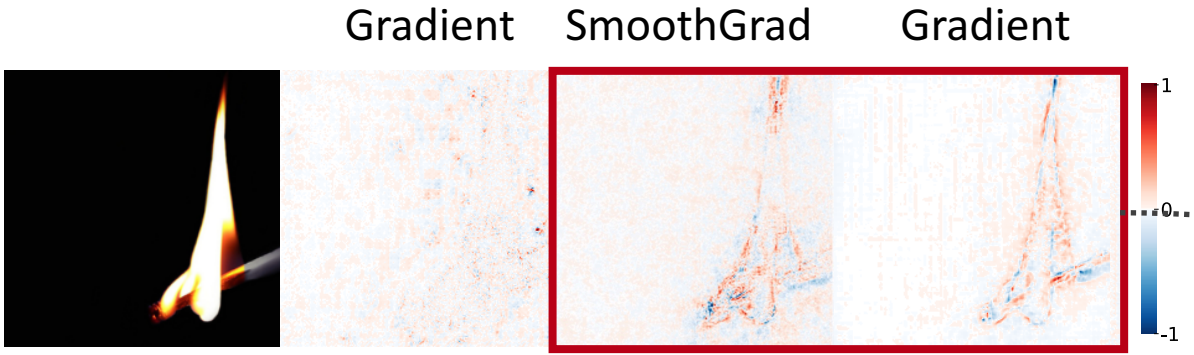
Gradient



GoogLeNet-R

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

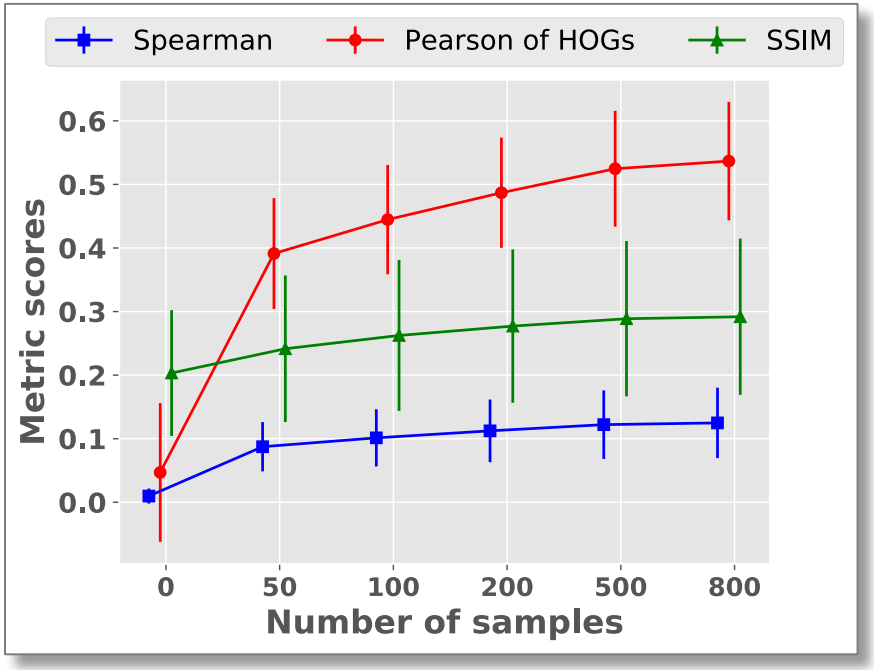
#2: Smoothed gradients can be misinterpreted



The **flame** is the most “important”

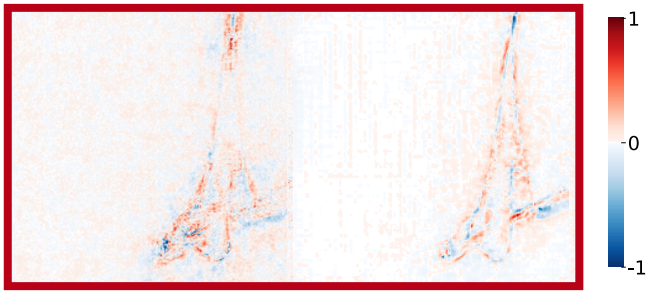
Some background pixels are the most “important”

#2: Smoothed gradients can be misinterpreted



Bansal*, Agarwal*, Nguyen*, 2020

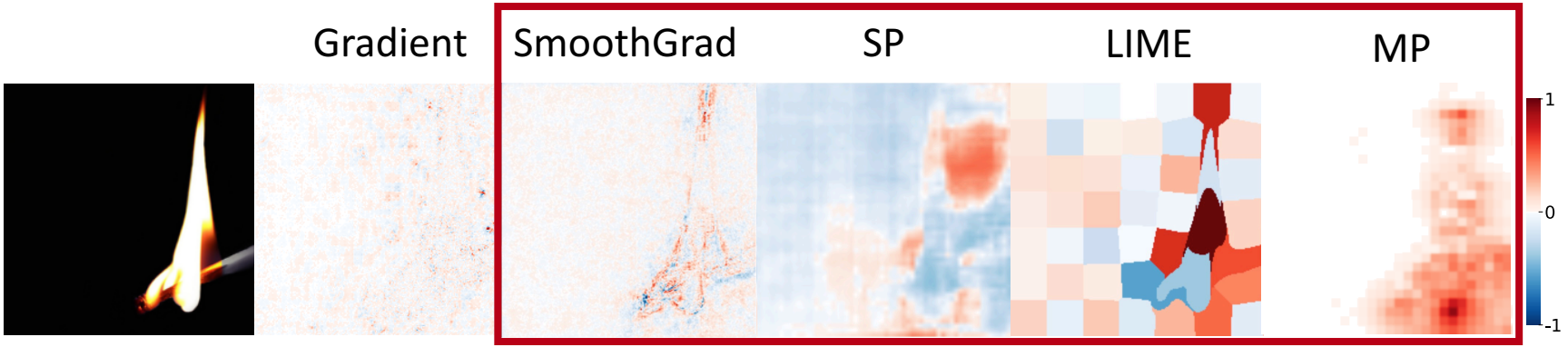
SmoothGrad Gradient



GoogLeNet ≠ **GoogLeNet-R**

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

#3: Many attribution maps are sensitive to hyperparams



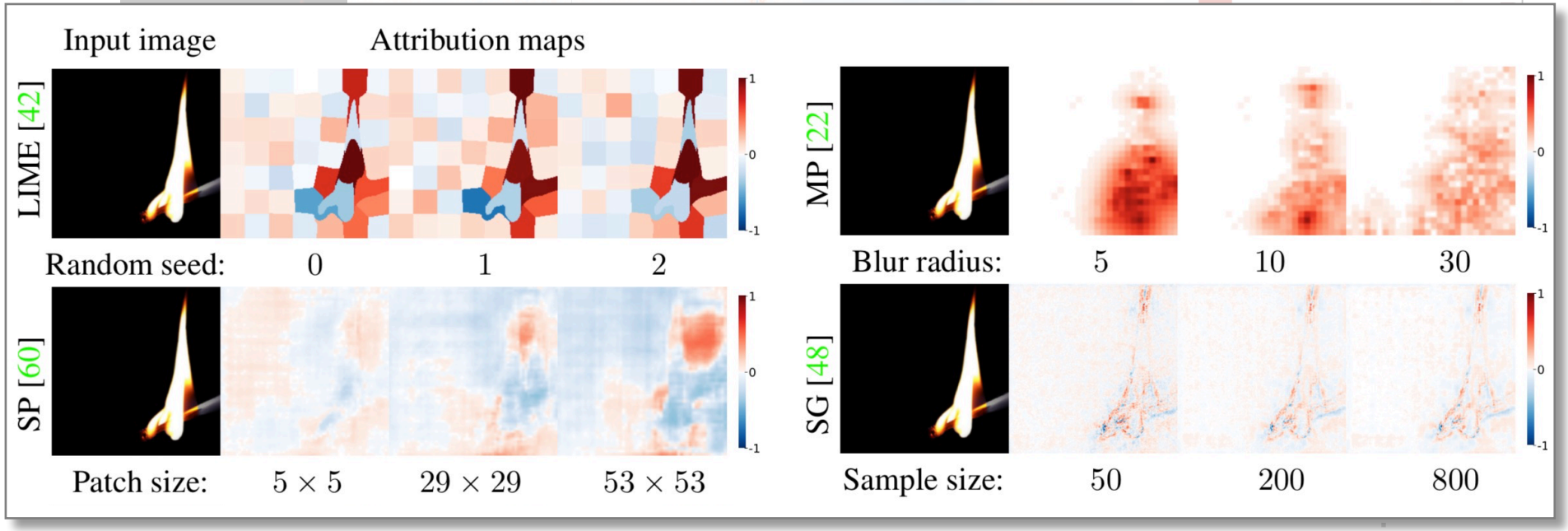
input model

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

hyperparameters

#3: Many attribution maps are sensitive to hyperparams

Gradient SmoothGrad SP LIME MP



hyperparameters

Bansal*, Agarwal*, Nguyen*, 2020

#3: Many attribution maps are sensitive to hyperparams

IN DEPTH | COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

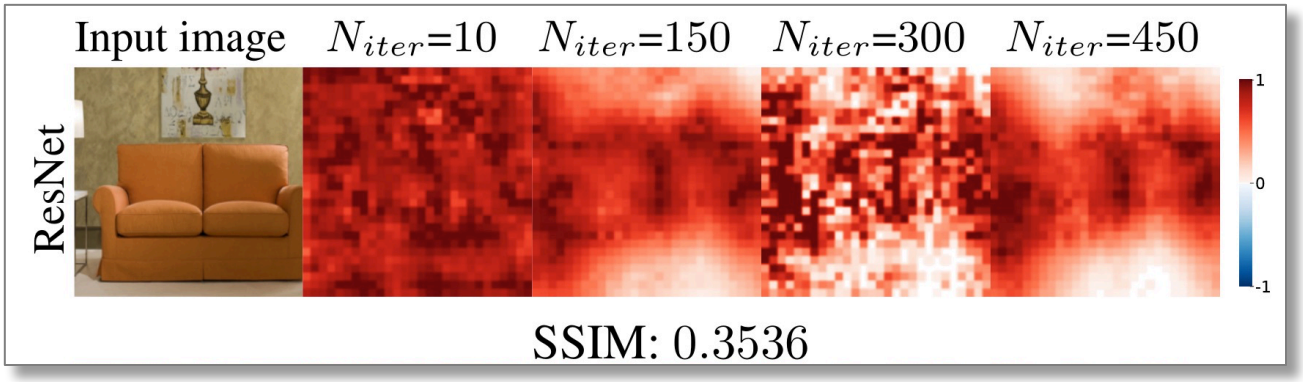
Matthew Hutson
+ See all authors and affiliations

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

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



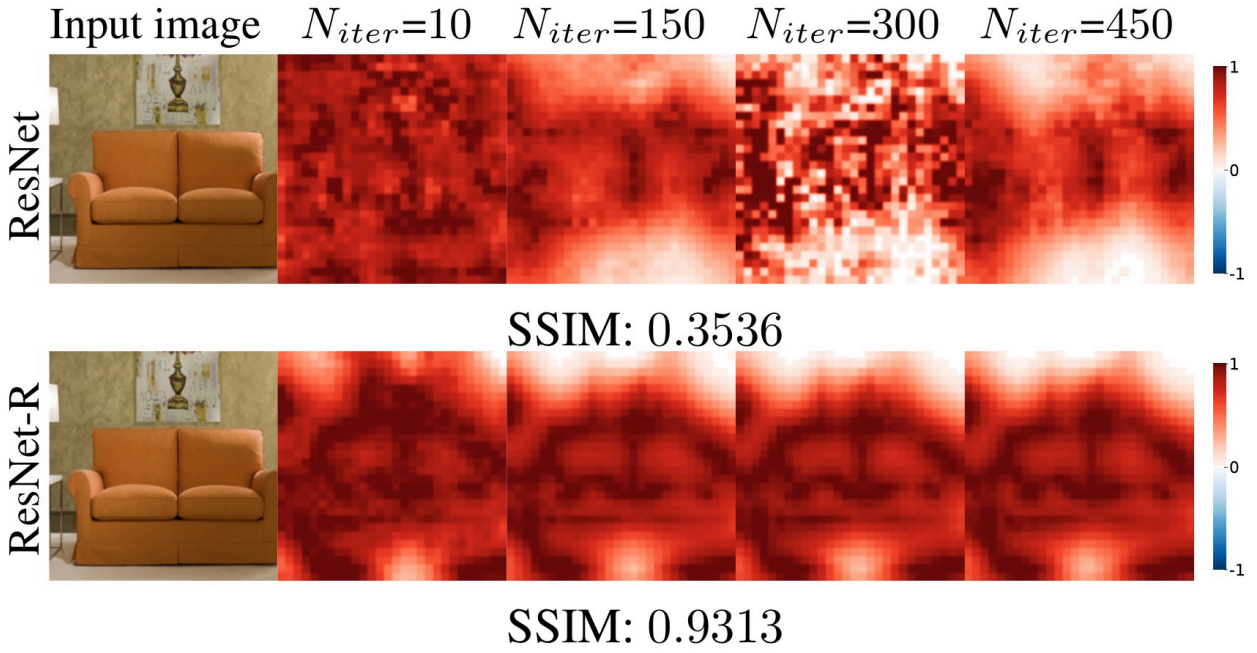
Idea: Identify a minimal region s.t. when blurred out would minimize classification score

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

Bansal*, Agarwal*, Nguyen*, 2020

Meaningful-Perturbation (MP)

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

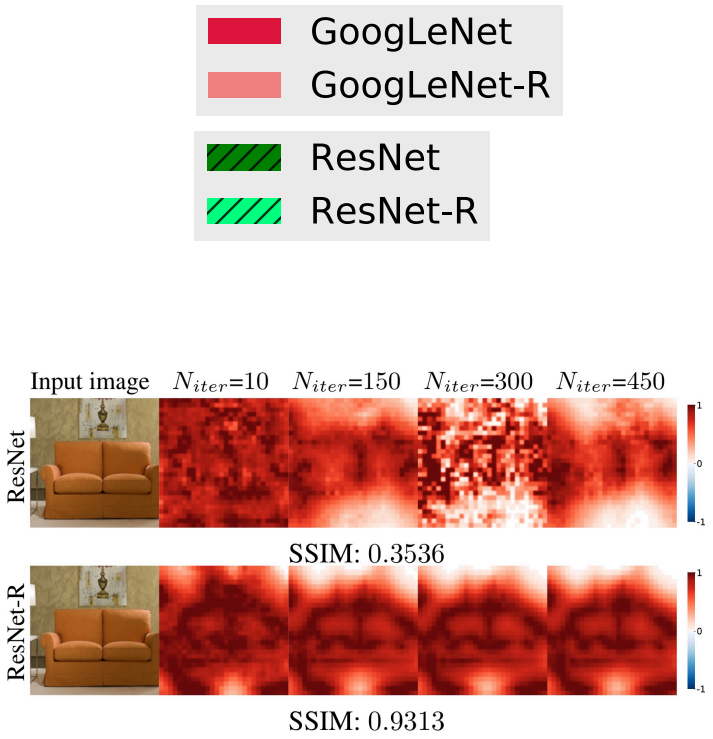


(b) Sensitivity to changes in the number of iterations N_{iter}

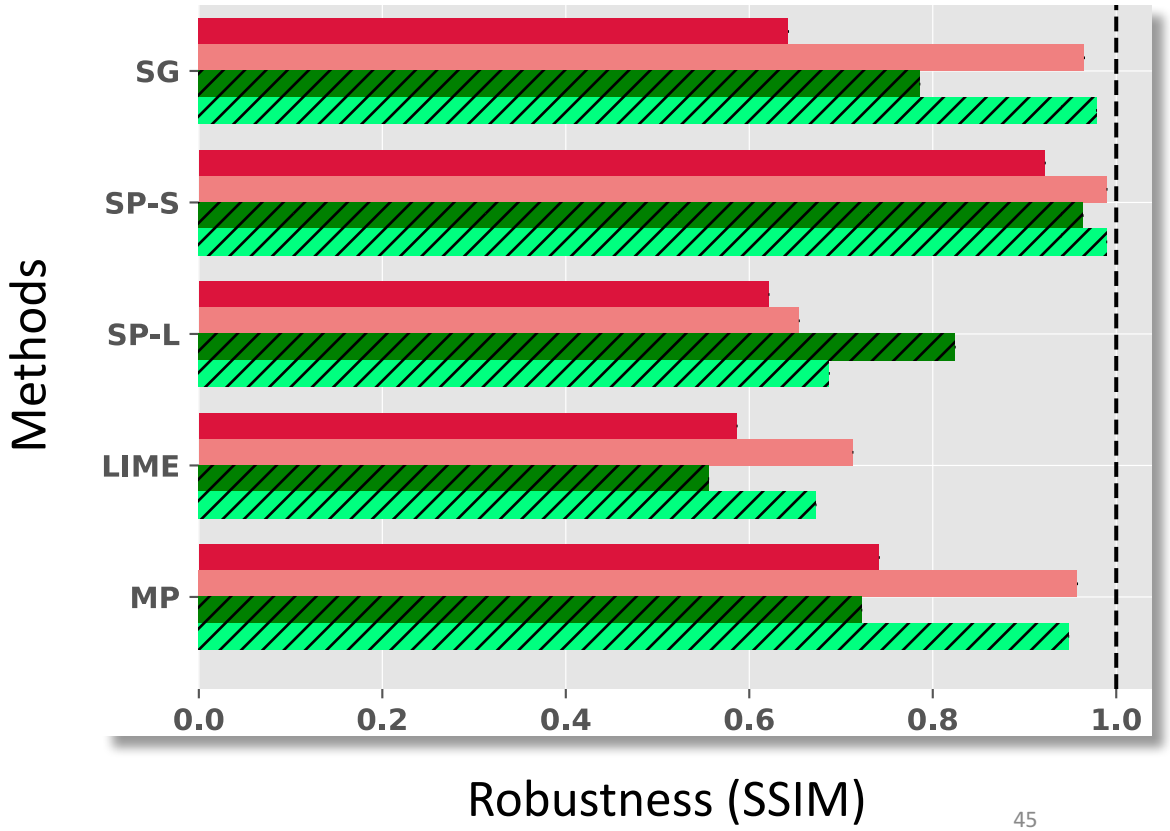
Bansal*, Agarwal*, Nguyen*, 2020

Meaningful-Perturbation (MP)

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



Bansal*, Agarwal*, Nguyen*, 2020



Conclusions

~~How-to~~

How to explain neural networks?


C Agarwal

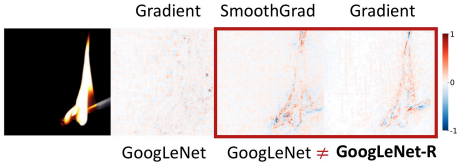

N Bansal

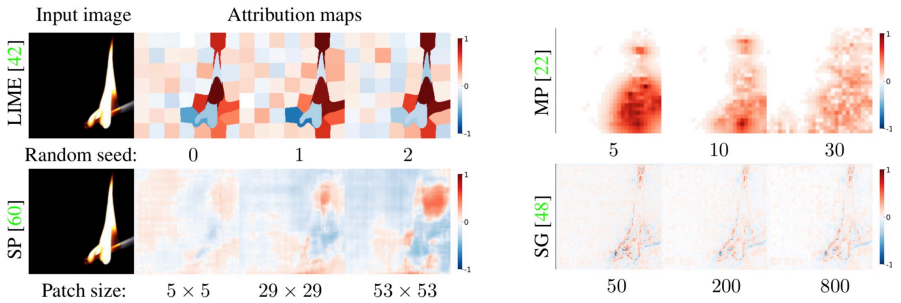

M Alcorn

1. Saliency maps for robust classifiers are not as noisy
2. Smoothed gradients can be misinterpreted
3. Many attribution methods are sensitive to hyper-parameters
4. For robust classifiers, attribution maps are more robust

Papers & code: http://AnhNguyen.me

< hiring PhD students! >





Work funded by