BOSTON UNIVERSITY

Deep Learning for Data Science DS 542

Lecture 19
Explanations

Images from cited papers.

Today’s Topic…

- Explanations
  - Why did our network make the decision that it did?
  - Will mostly talk about image explanations, but mostly transferable.

Two High Level Approaches to Explanations

High level techniques:

- Gradient Analysis
- Shapley Values

Mostly focused on gradient-based techniques today.

Same underlying intuitions:

- If I change X, what happens?
- Which present Xs drove this classification?

Gradients

General idea:

- What inputs have highest gradients for our output?
- If classifying inputs, which inputs could change classification result fastest?

Salience

Dictionary

Definitions from Oxford Languages · Learn more

sa·li·ence

/ˈsālyəns/

noun

the quality of being particularly noticeable or important; prominence.
"the political salience of religion has a considerable impact"

What Image Best Represents a Class?

Given a neural network, what input maximizes each class output?

- Is this the best way to understand what the network is looking for?
- This is not specific to a any particular image.
- But we can calculate this for any pre-trained neural network.

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2014)

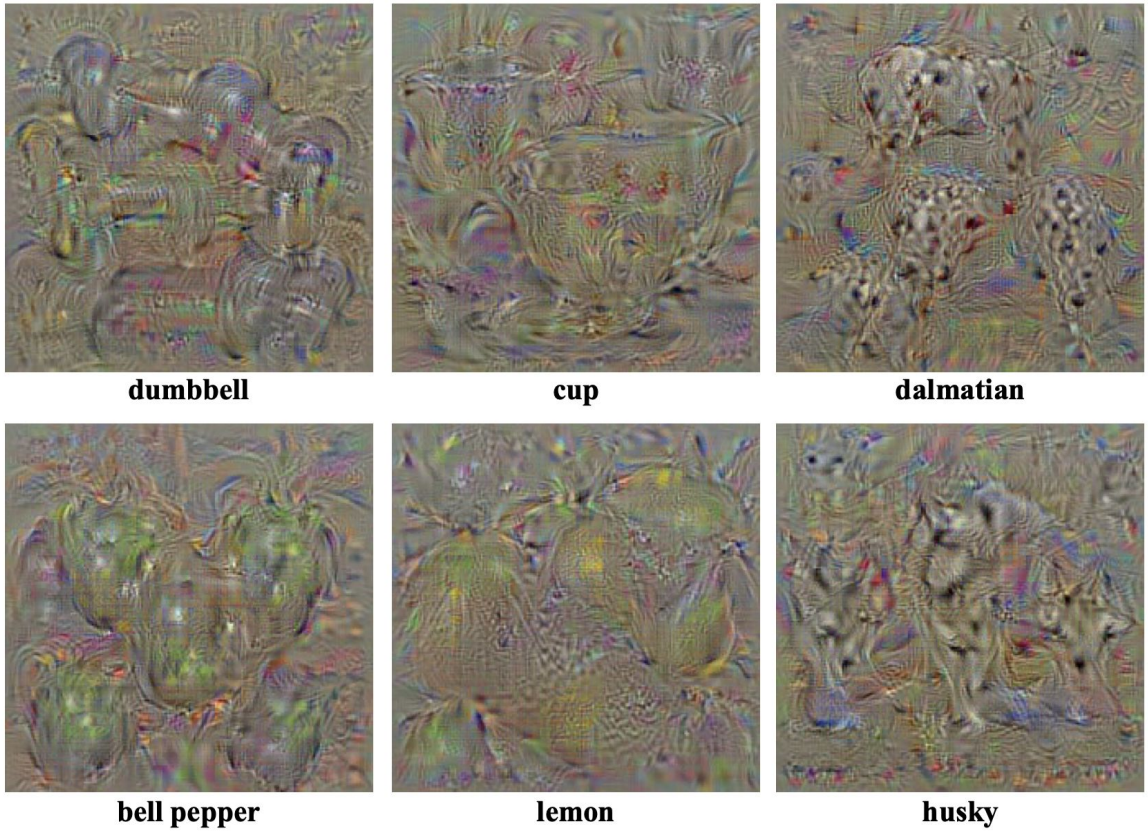
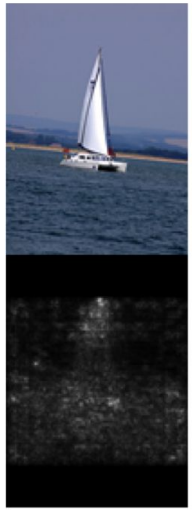
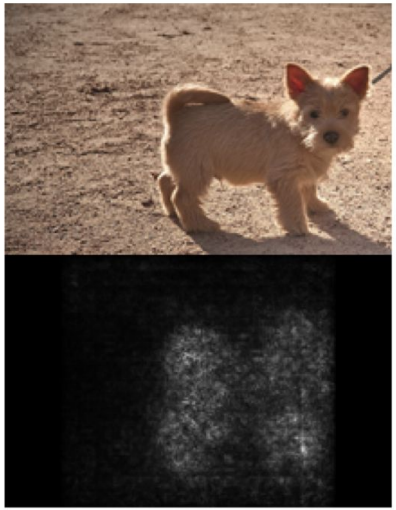
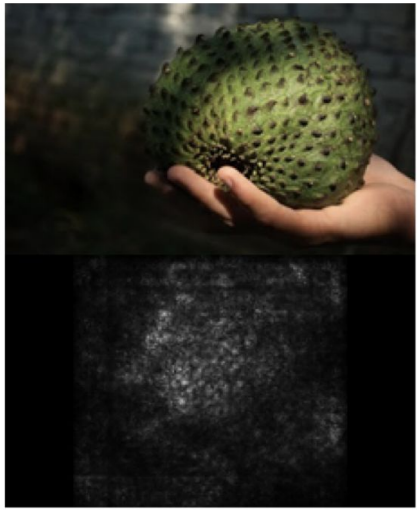


Image-Specific Class Saliency Visualisation

“Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images. The maps were extracted using a single back-propagation pass through a classification ConvNet. No additional annotation (except for the image labels) was used in training.”







Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2014)

Image-Specific Class Saliency Visualisation

“Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images. The maps were extracted using a single back-propagation pass through a classification ConvNet. No additional annotation (except for the image labels) was used in training.”

[Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps](#)(2014)

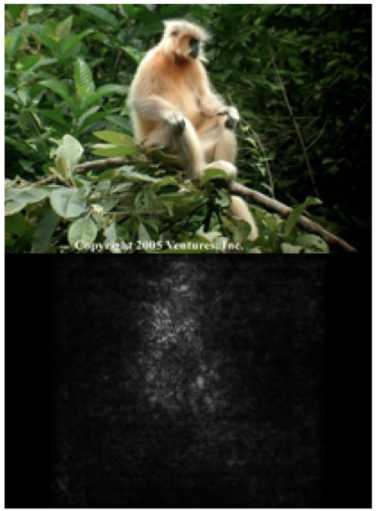


Image-Specific Class Saliency Visualisation

“Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images. The maps were extracted using a single back-propagation pass through a classification ConvNet. No additional annotation (except for the image labels) was used in training.”

[Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps](#) (2014)

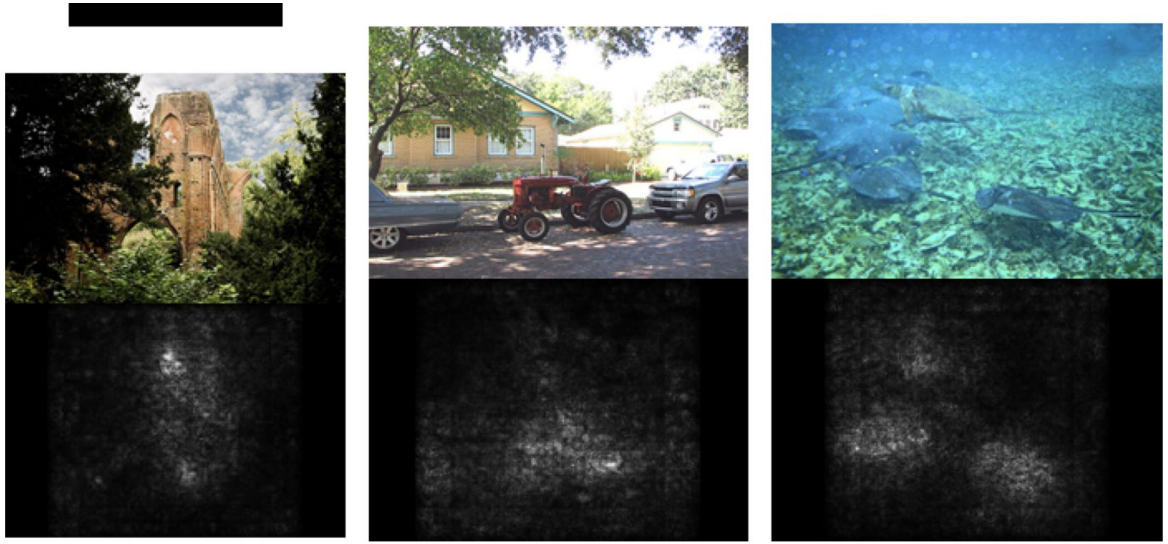


Image-Specific Class Saliency Visualisation

How did those saliency maps work?

Linear example for image I and class c:

$$S_c(I) = w_c^T I + b_c,$$

Linear approximation using gradients:

$$S_c(I) \approx w^T I + b, \quad w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$

Note: this includes per-pixel gradients for images.

Image-Specific Class Saliency Visualisation

How to map this to a saliency map?

$$S_c(I) \approx w^T I + b,$$

For grayscale images, take absolute value of pixel gradient.

$$M_{ij} = |w_{h(i,j)}|,$$

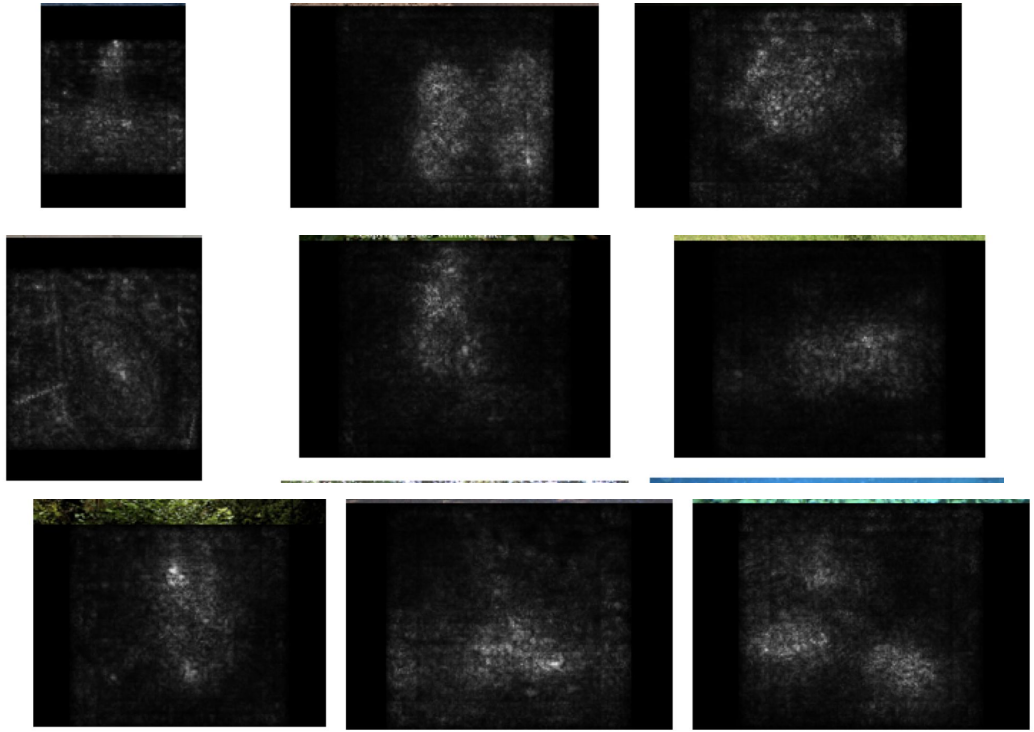
For color images, take max absolute value over all color channels.

$$M_{ij} = \max_c |w_{h(i,j,c)}|.$$

Image-Specific Class Saliency Visualisation

Simple concept -

- Basically just use class output gradients w/respect to inputs.
- But pretty fuzzy output.
- Do you remember what these were?



SmoothGrad

Basic idea:

- Rerun the same gradient-based process adding noise each time.
  - Then average the results together to get a smoothed version.
  - Specifically motivated by noisy gradients.
- Note: this paper uses the phrase “sensitivity map” instead of “saliency map”.

[SmoothGrad: removing noise by adding noise](#) (2017)

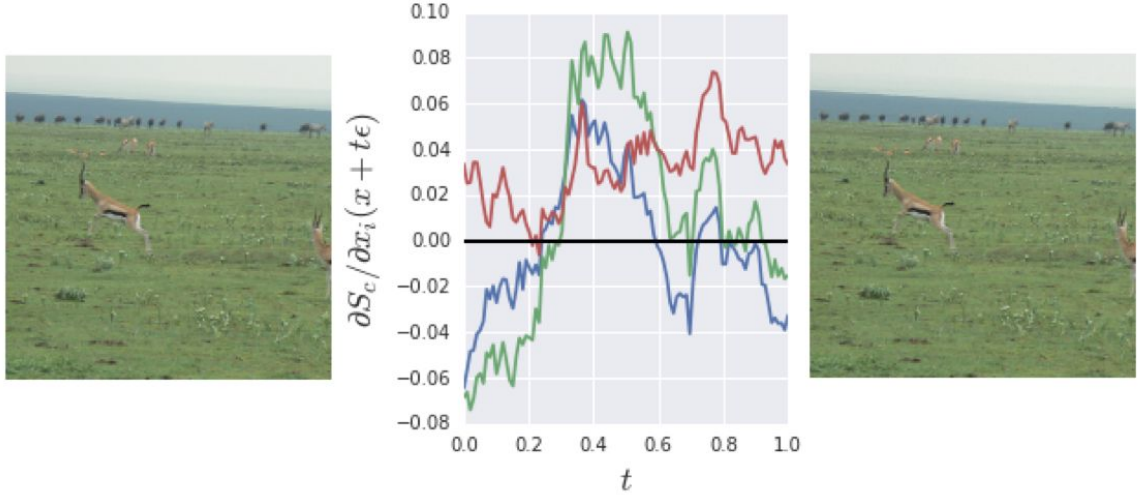
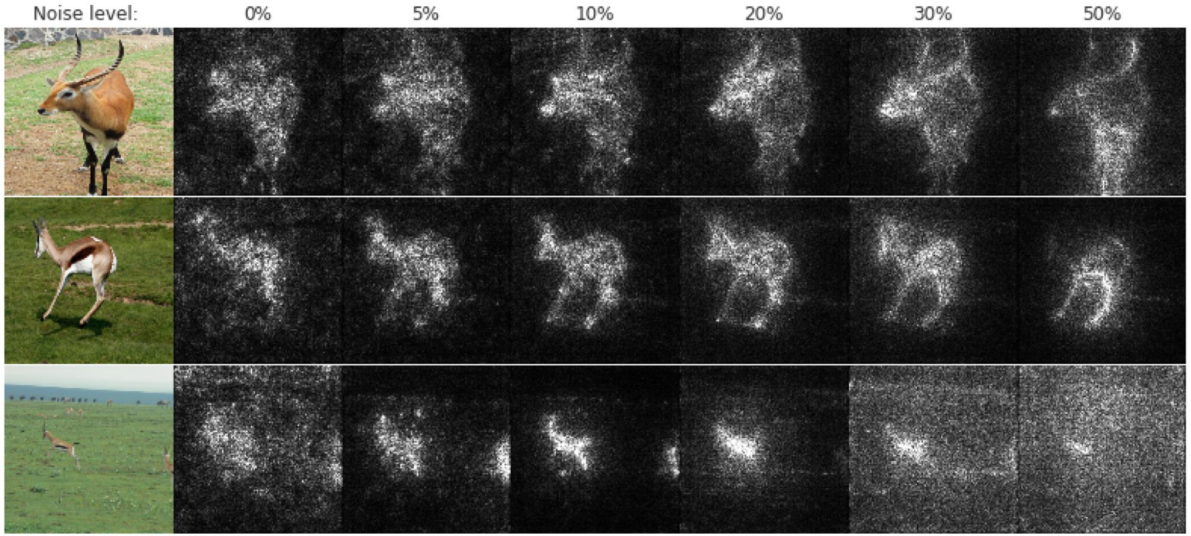


Figure 2. The partial derivative of S_c with respect to the RGB values of a single pixel as a fraction of the maximum entry in the gradient vector, $\max_i \frac{\partial S_c}{\partial x_i}(t)$, (middle plot) as one slowly moves away from a baseline image x (left plot) to a fixed location $x + \epsilon$ (right plot). ϵ is one random sample from $\mathcal{N}(0, 0.01^2)$. The final image $(x + \epsilon)$ is indistinguishable to a human from the origin image x.

SmoothGrad

Average of 50 trials adding Gaussian noise and computing saliency…

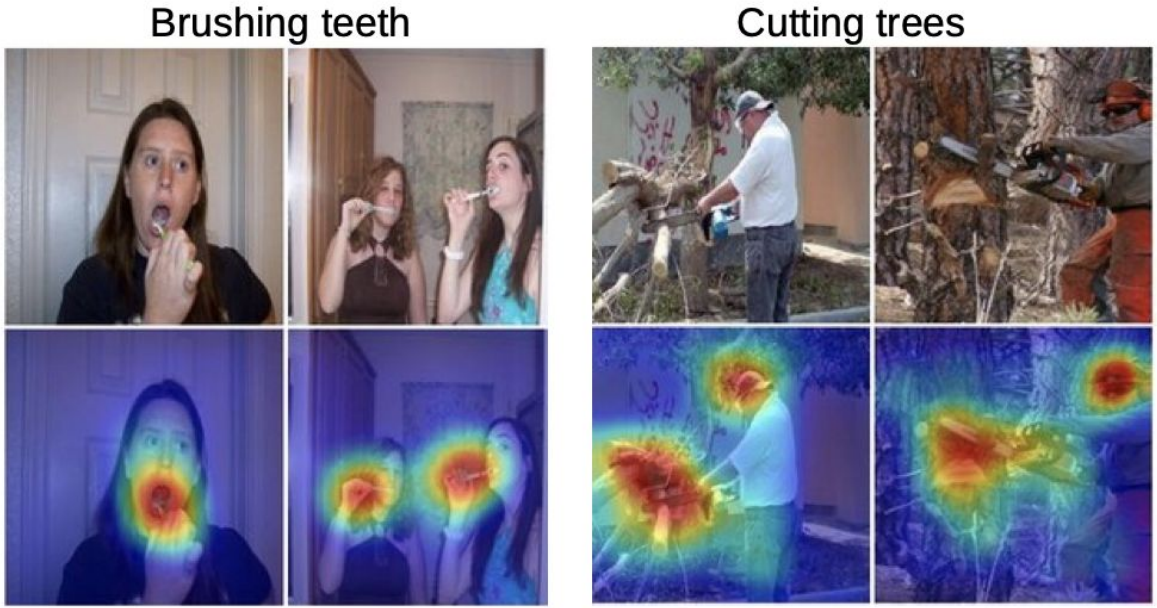
- Noise level = $\sigma / (x_{max} - x_{min})$
- No universal best noise level?



SmoothGrad: removing noise by adding noise (2017)

Class Activation Maps

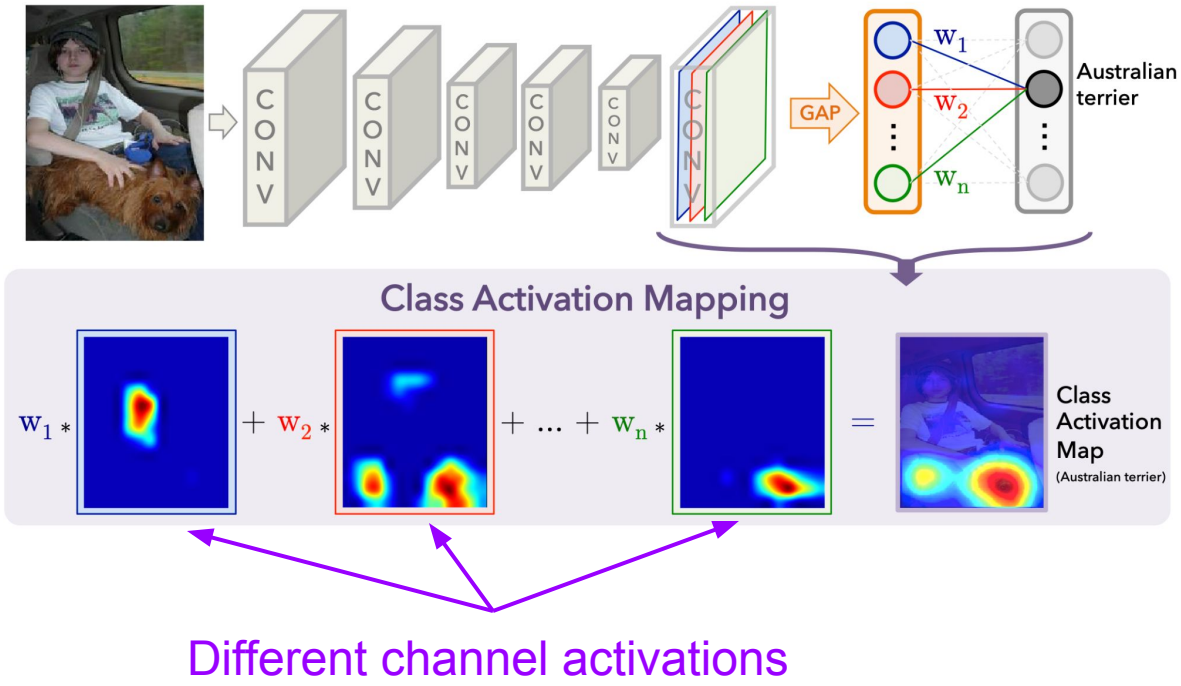
- Trace class outputs back to last convolutional layer.
- Last convolutional layer has (low res) positional information



Learning Deep Features for Discriminative Localization (2015)

Class Activation Maps

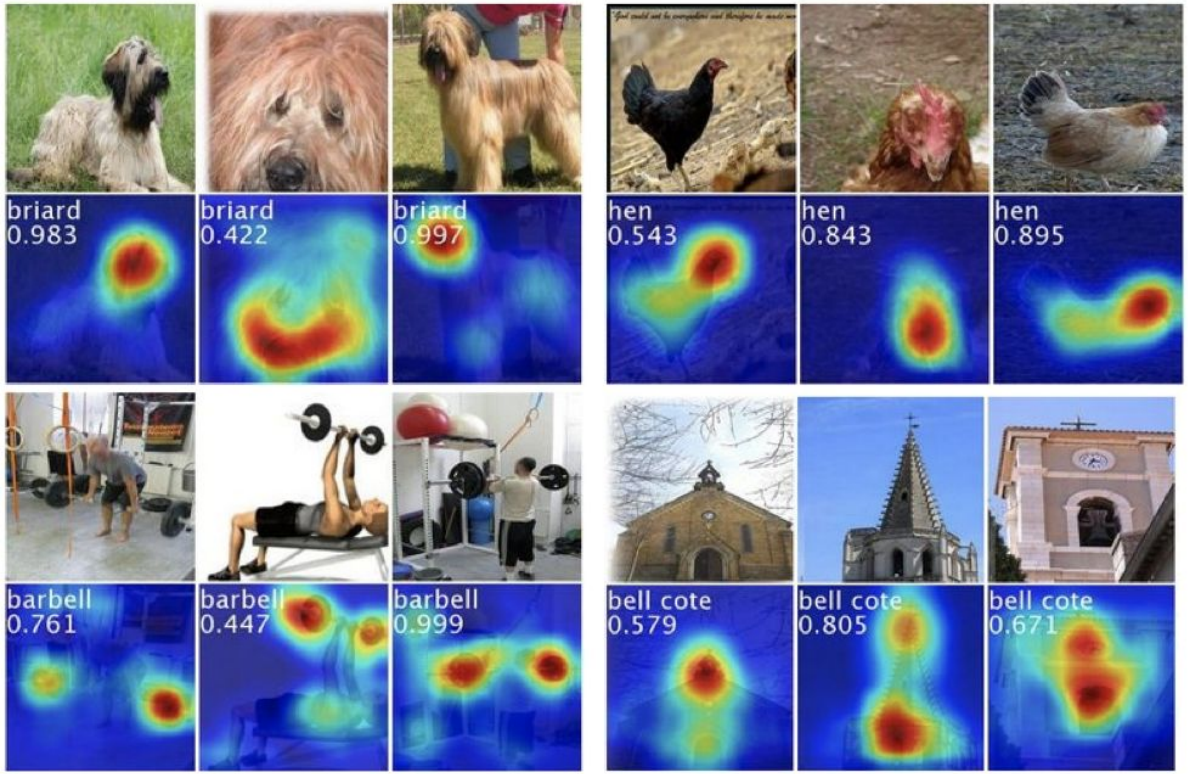
- Global average pooling layers aggregate each convolutional channel.
- Then weighted linear combination for each class.
- Class activation mapping = convolution activation * class weight



[Learning Deep Features for Discriminative Localization](#) (2015)

Class Activation Maps

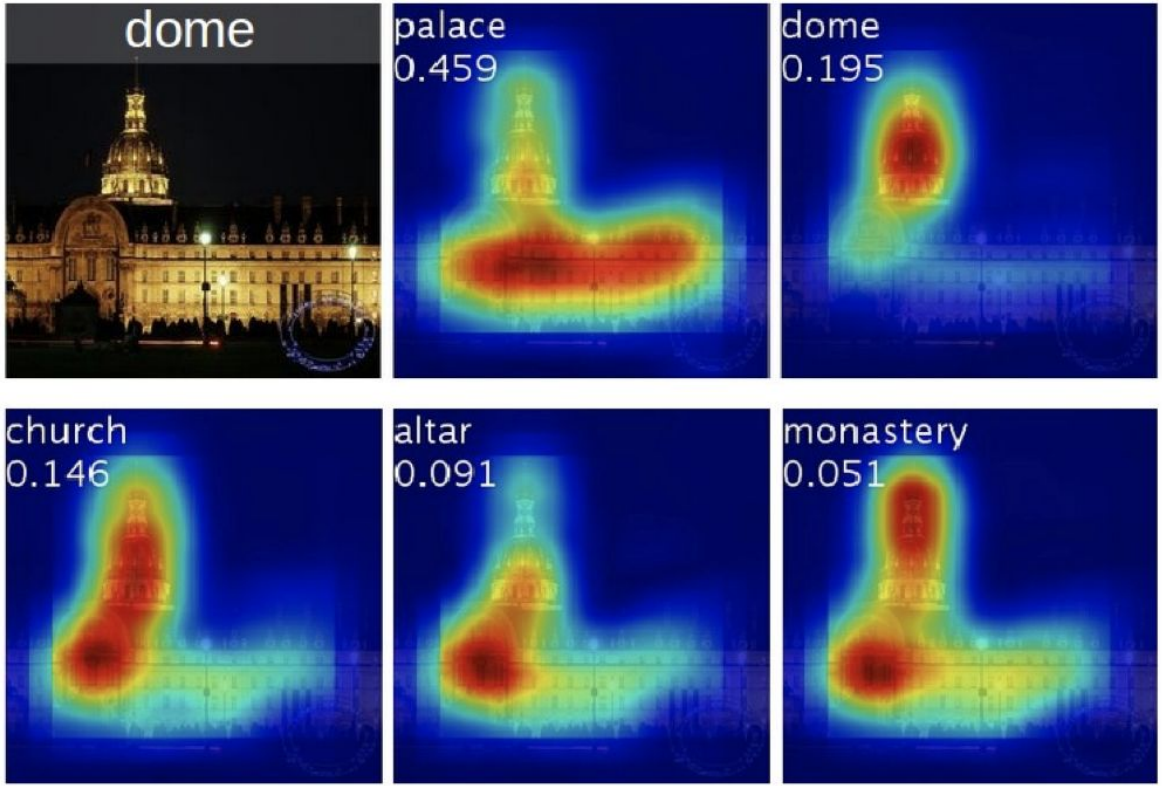
- Class activation maps appear to focus on particular parts of animals and other items.
- Smoothing effects from last convolutional not being at full resolution.



Learning Deep Features for Discriminative Localization (2015)

Class Activation Maps

- Class activation maps are weighted averages of the last convolution channel activations.
- They can take different shapes because different classes have different weights.
- But similar areas are noticeable, especially across related classes.

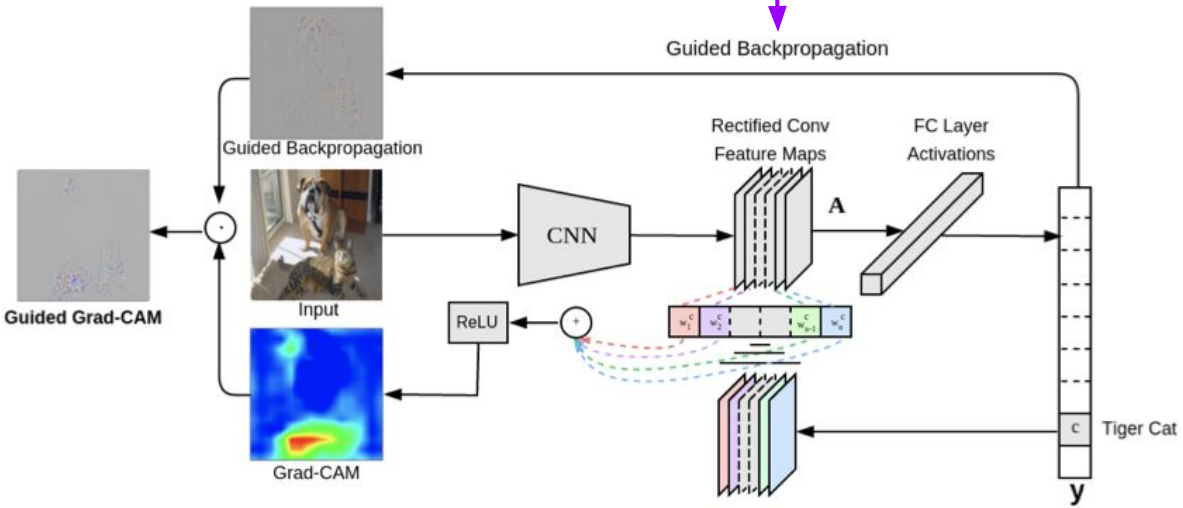


Learning Deep Features for Discriminative Localization (2015)

Gradient-weighted Class Activation Mapping

- Generalize CAM to deep networks
- Backpropagate gradients for a particular class output.
- Pick a particular convolutional layer for GRAD-CAM.

Heavily filtered backpropagation variant, used for per-pixel gradients.



[Grad-CAM: Why did you say that?](#) (2017)

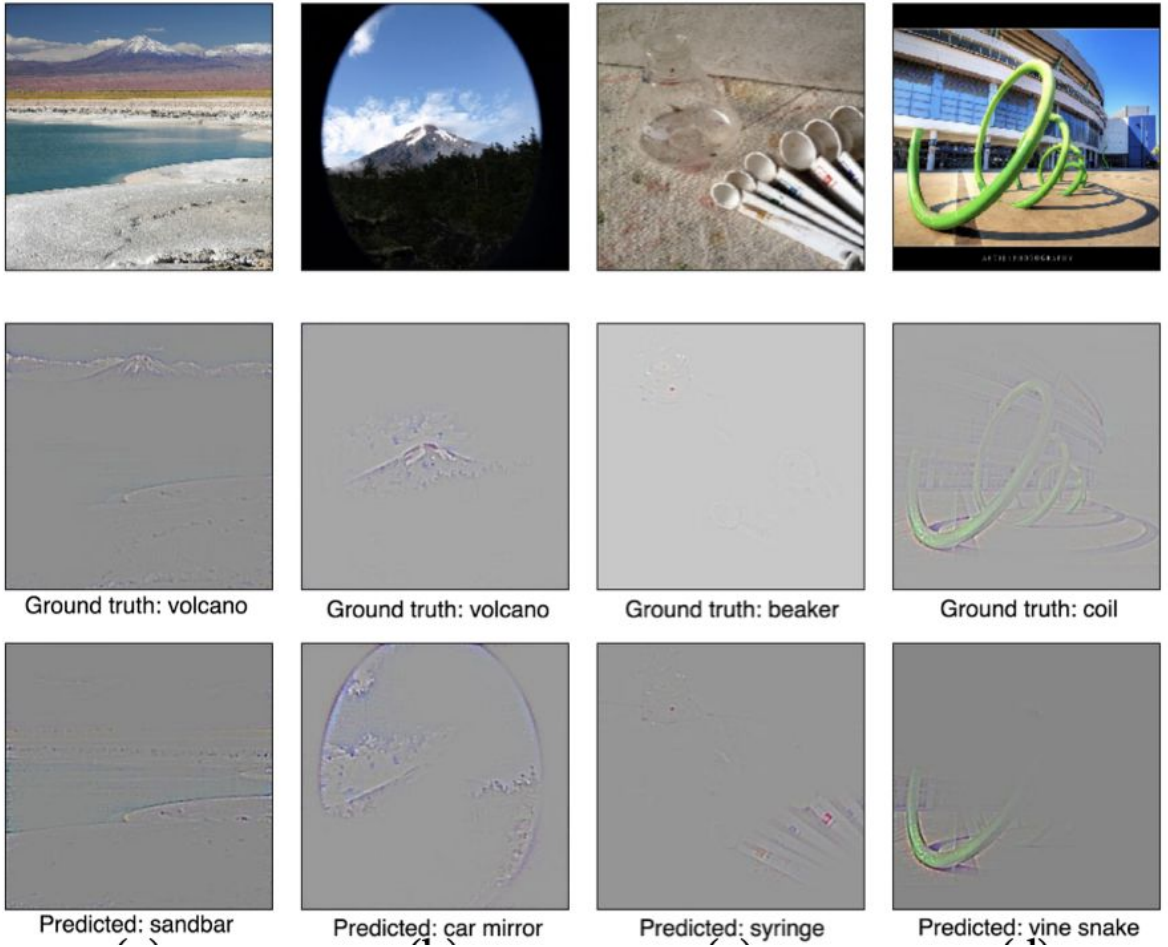
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2019)

Gradient-weighted Class Activation Mapping

- Very different output - a combination of the heat map and original image.
- These examples look at cases where the model made a mistake…

[Grad-CAM: Why did you say that?](#) (2017)

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2019)



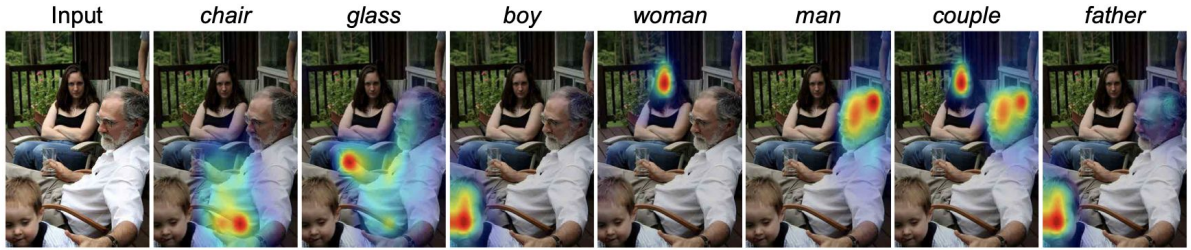
Beware - Gradient Maps Can Be Misleading!

“One network has randomly initialized weights, the other gets >99% accuracy on the test set.”

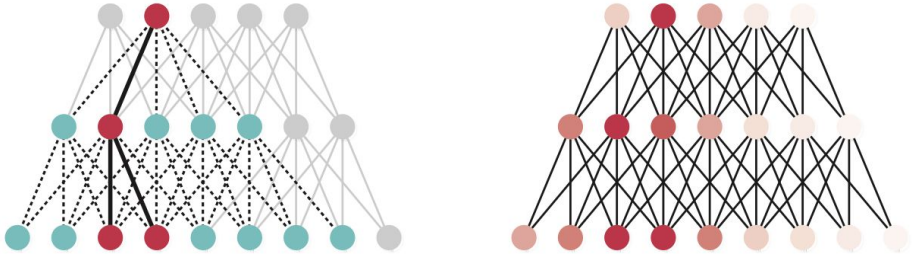
Visualizing the Impact of Feature Attribution Baselines (2020)

Excitation Backprop

- Not the same attention that we've talked about recently.
- Use a sampling process to try to identify the most relevant input pixels via back propagation.
- More sophisticated analysis than previous deterministic versions.



[Top-down Neural Attention by Excitation Backprop](#) (2016)



(a) Deterministic WTA (b) Probabilistic WTA (c) Winner Sampling

Excitation Backprop

- Can get different maps at different resolutions from different layers…

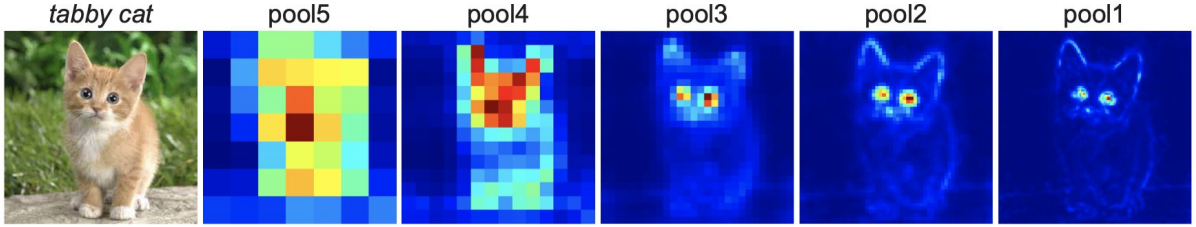


Fig. 3. Example Marginal Winning Probability (MWP) maps computed via Excitation Backprop from different layers of the public `VGG16` model [29] trained on ImageNet. The input image is shown on the right. The MWP maps are generated for the category `tabby cat`. Neurons at higher-level layers have larger receptive fields and strides. Thus, they can capture larger areas but with lower spatial accuracy. Neurons at lower layers tend to more precisely localize features at smaller scale.

Top-down Neural Attention by Excitation Backprop (2016)

Excitation Backprop

- A variation on contrastive techniques improved performance.

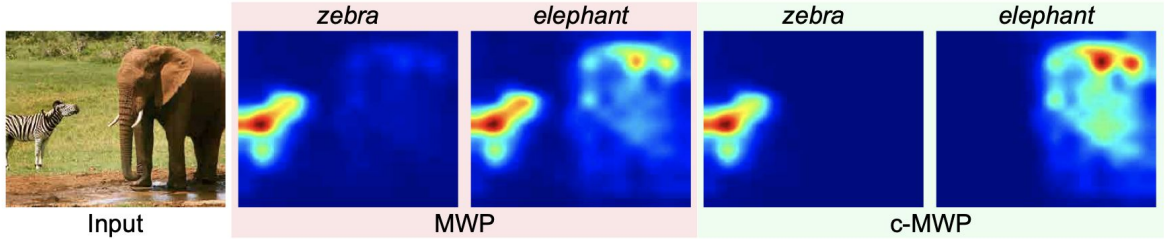


Fig. 4. Marginal Winning Probability (MWP) *vs.* contrastive MWP (c-MWP). The input image is resized to 224×224, and we use `GoogleNet` pretrained on ImageNet to generate the MWP maps and c-MWP maps for `zebra` and `elephant`. The MWP map for `elephant` does not successfully suppress the zebra. In contrast, by cancelling out common winner neurons for `elephant` and `non-elephant`, the c-MWP map more effectively highlights the elephant.

Top-down Neural Attention by Excitation Backprop (2016)

Evaluation: Pointing Game

Given attention mechanism of excitation backprop, can sample most relevant points according to the model…

- How many of them match a human-labeled ground truth?
- Score on accuracy: hits / (hits + misses)

Don’t need attention for this

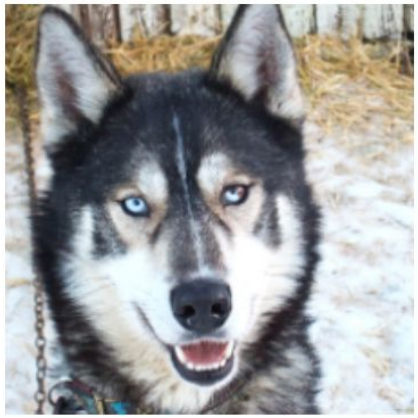
- Use the most salient pixels?
- Or sample pixels by saliency?

LIME

Main ideas:

- Make local approximations of model that are explainable.
- Use linear models as easy explainable model.
- Use “super pixels” as better modeling chunk for images.

“Why Should I Trust You?” Explaining the Predictions of Any Classifier (2016)



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

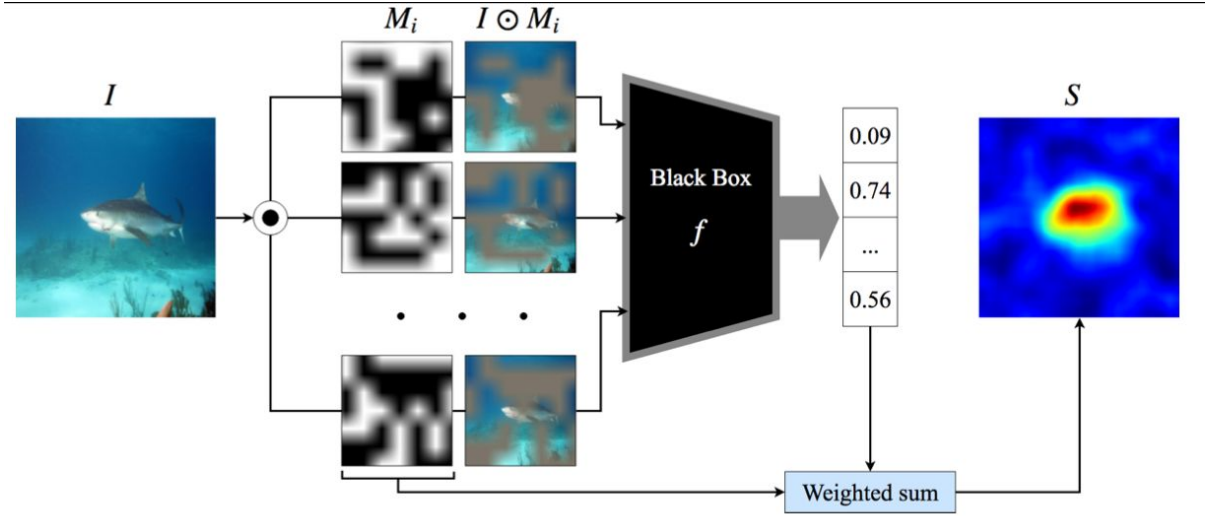
Table 2: “Husky vs Wolf” experiment results.

RISE

Key idea:

- Repeatedly mask image to see whether classifier still can make the same classification.

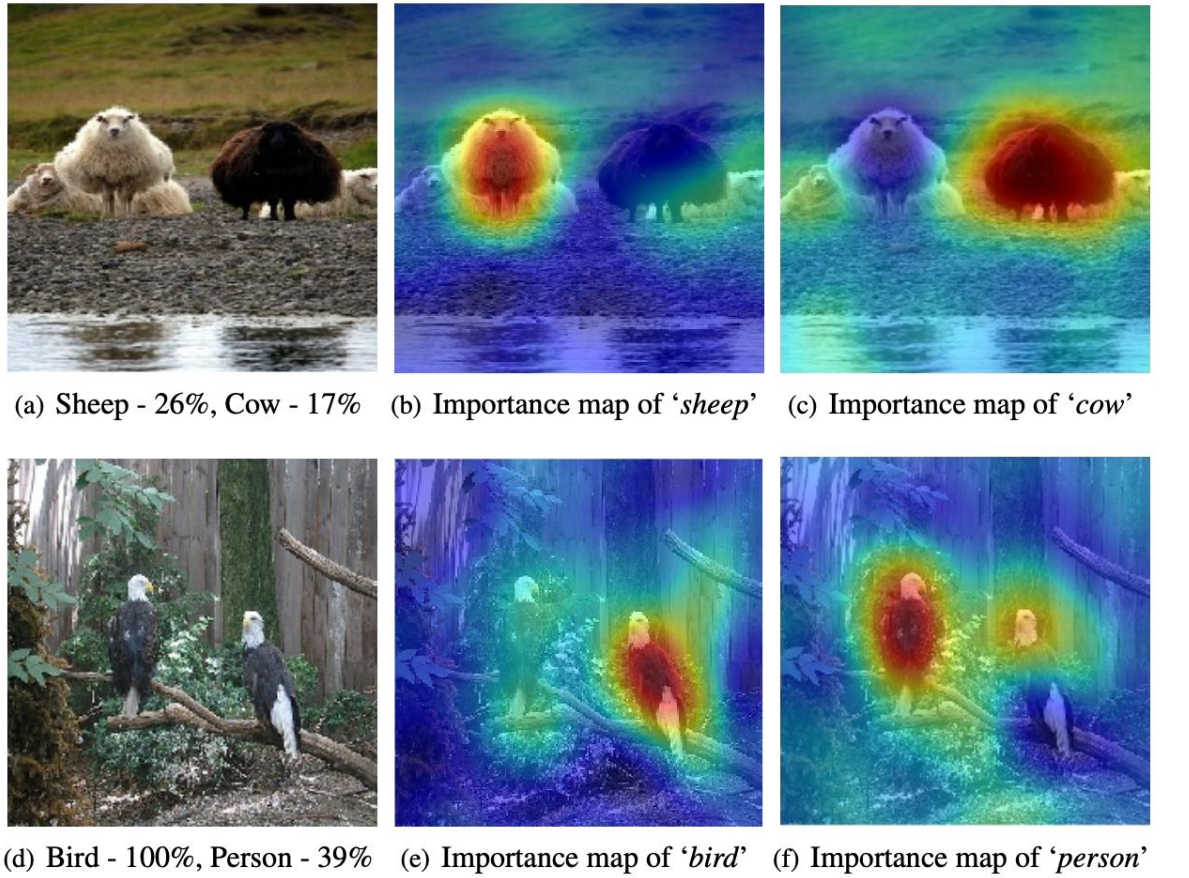
[RISE: Randomized Input Sampling for Explanation of Black-box Models](#)
(2018)



RISE

- Weighting masks by classification output emphasizes areas required for classification.

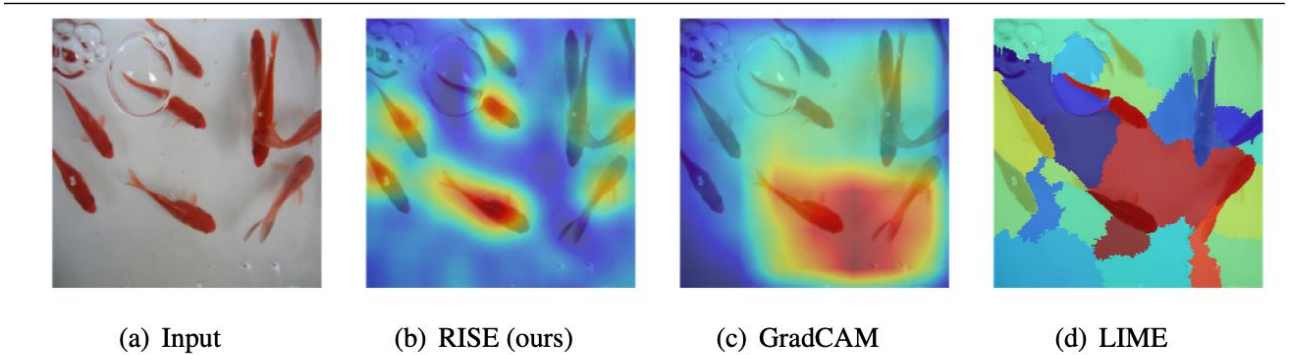
[RISE: Randomized Input Sampling for Explanation of Black-box Models](#) (2018)



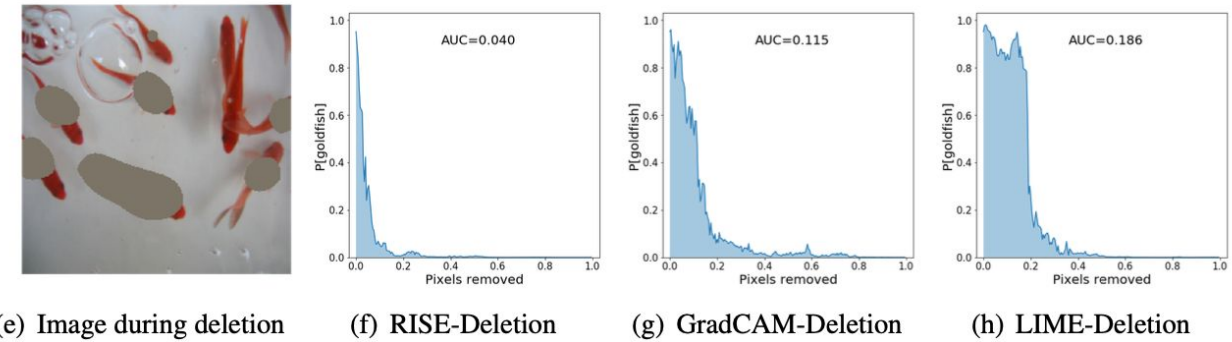
Evaluation: Deletion

How fast does the classification probability drop if you “delete” important pixels?

- Score by area under the curve.
- Lower is better.



(a) Input (b) RISE (ours) (c) GradCAM (d) LIME



(e) Image during deletion (f) RISE-Deletion (g) GradCAM-Deletion (h) LIME-Deletion

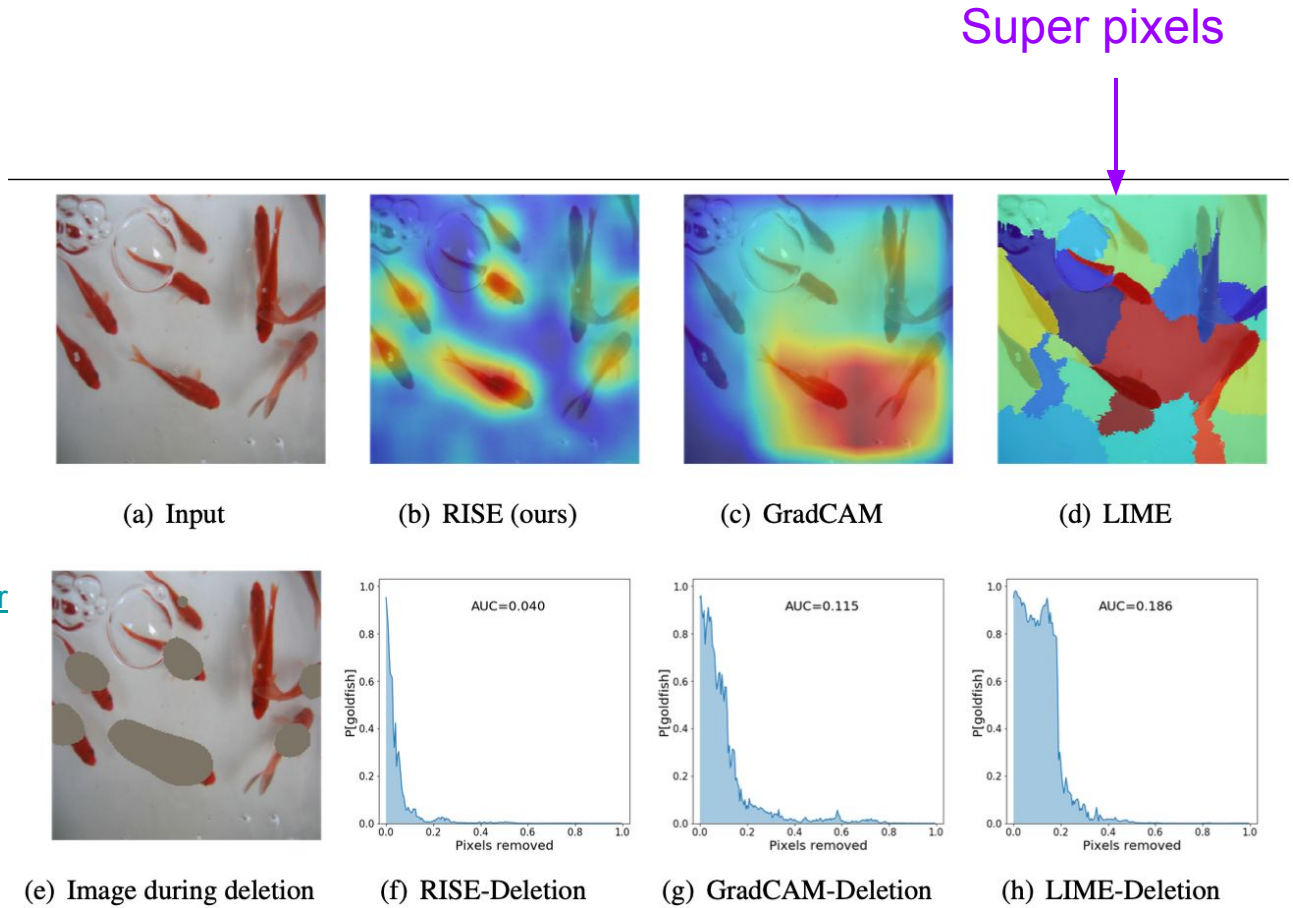
[RISE: Randomized Input Sampling for Explanation of Black-box Models](#) (2018)

Evaluation: Deletion

How fast does the classification probability drop if you “delete” important pixels?

- Score by area under the curve.
- Lower is better.

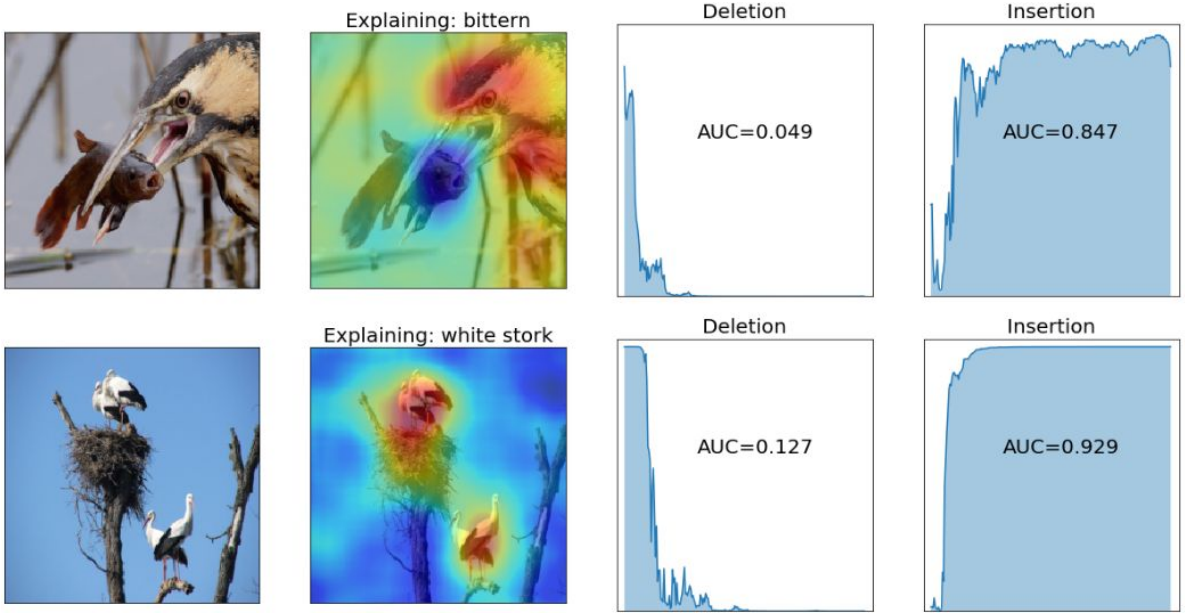
[RISE: Randomized Input Sampling for Explanation of Black-box Models](#) (2018)



Evaluation: Insertion

How fast does the classification probability increase if you “insert” important pixels?

- Flipped version of deletion.
- Score by area under the curve.
- Higher is better.



[RISE: Randomized Input Sampling for Explanation of Black-box Models](#) (2018)

Feedback?

Shapley Values

Shapley Additive Explanations

A Unified Approach to Interpreting Model Predictions (2017)