Interpretability Tutorial

Ruth Fong

University of Oxford, VGG Reading Group

Tuesday, 8 October 2019

What do we mean when we say “interpretability”?

When is interpretability desirable?

When is interpretability desirable?

1. More emphasis in sensitive domains
2. Less emphasis given historical performance
3. Less emphasis if improving interpretability incurs other costs

What are desiderata for interpretability research?

Desiderata for Interpretability Research

1. Trust
   - still subjective
   - “well-understood” or “confidence-giving”
   - “how often a model is right” + “for which examples is it right”
2. Causality
   - “does the model learn *causal* relations?”
3. Transferability
   - “does the model *generalize*?”
4. Informativeness
   - “what information does can this model provide to human decision makers?”
5. Fair and Ethical Decision Making
   - “can we produce *interpretations* by which to assess if automated decisions conform to ethical standards?”
   - “right to explanation” — EU GDPR

[Lipton, ICML WHI 2016]

What has interpretability research focused on thus far?

See white paper.

1. Explaining a specific prediction

- “local explanation”
- What part of the input is responsible for the model’s prediction? **Attribution maps.**
- Which training examples are are responsible for the model’s prediction? **Attribution to data points.**
- Disadvantage: Only explains local behavior around a a given “point”

2. Explaining global model behavior

- “transparency”
- Can we construct *human-understandable* representations of the model’s global behavior? **Model {distillation,approximation, compression}.**
- What properties does a component of the model have? How does a component of a model functionally work? **“Science of X.”**

3. Building more interpretable models

- Related to model approximation
- Disadvantage: No consensus yet on appropriate "interpretable" models

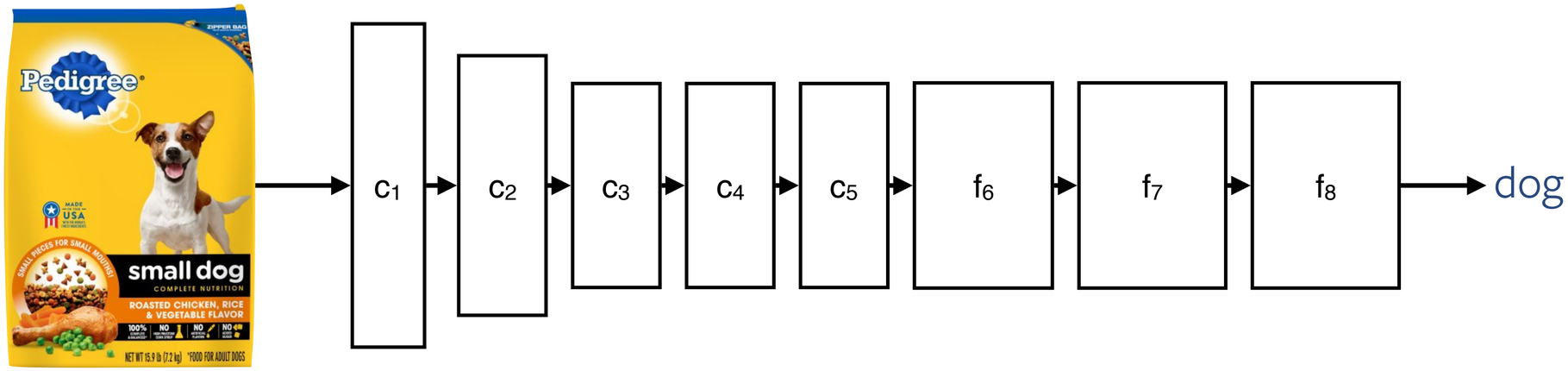
4. Visualization tools

- Static
  - input/output/intermediate representation
  - e.g, t-SNE, feature visualization
- Interactive
  - dashboards, explanations, interaction with model

Case study: Attribution maps
Evolution of methods and benchmarks

Attribution

Where is the model **looking**?



?

dog

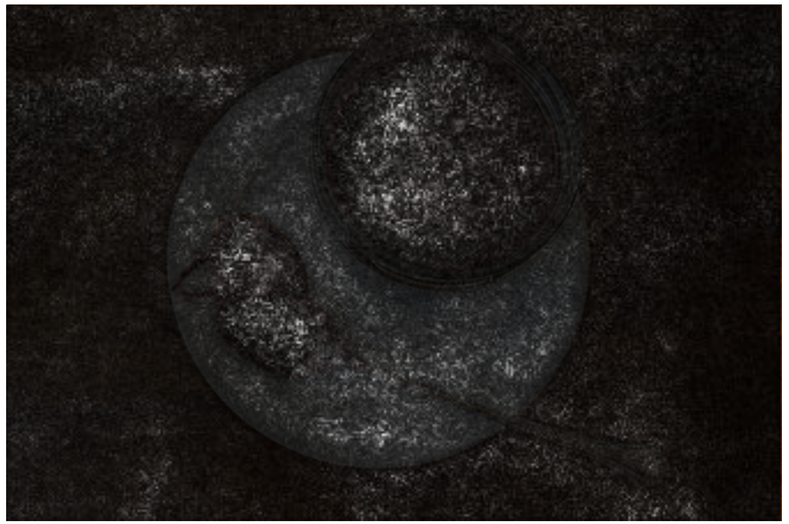
Popular Methods

Backpropagation

Combine network activations and gradients

Input

Gradient



Grad-CAM

Fast, but difficult to characterize

[Simonyan et al., ICLR Workshop 2014; Selvaraju et al., ICCV 2017]
[Mahendran and Vedaldi, ECCV 2016; Adebayo et al., NeurIPS 2018]

Backprop-based Methods: Improving the Gradient

$$\max \frac{df_c(x)}{dx}$$

1. Gradient
[Simonyan et al., 2014]

```
def relu_backward(x, dx):
  return (x > 0) * dx
```

2. DeConvNet
[Zeiler & Fergus, 2014]

```
def relu_backward(x, dx):
  return (dx > 0) * dx
```

3. Guided Backprop
[Springenberg et al., 2015]

```
def relu_backward(x, dx):
  return ((dx > 0) * (x > 0) *
  dx)
```

Evolution to Improve **Visual Quality**

Backdrop-based Methods: Mitigate “gradient saturation”

Gradient * Input

$$x \odot \frac{df_c(x)}{dx}$$

Integrated Gradients
[Sundararajan et al., 2017]

$$\hat{S}_c(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S_c(\bar{x} - \alpha(x - \bar{x})}{\partial x} d\alpha$$

* integrate over different intensities

* requires choice of $\bar{x}$

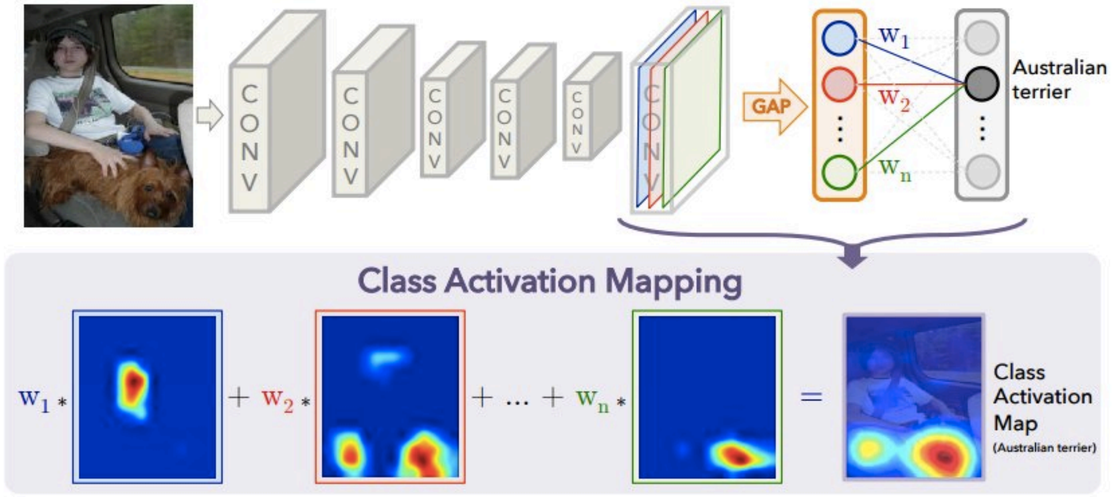
SmoothGrad [Smilkov et al., 2017]

$$\hat{S}_c(x) = \frac{1}{n} \sum^{n} S_c(x + \mathcal{N}(0, \sigma^2))$$

* average away the noise

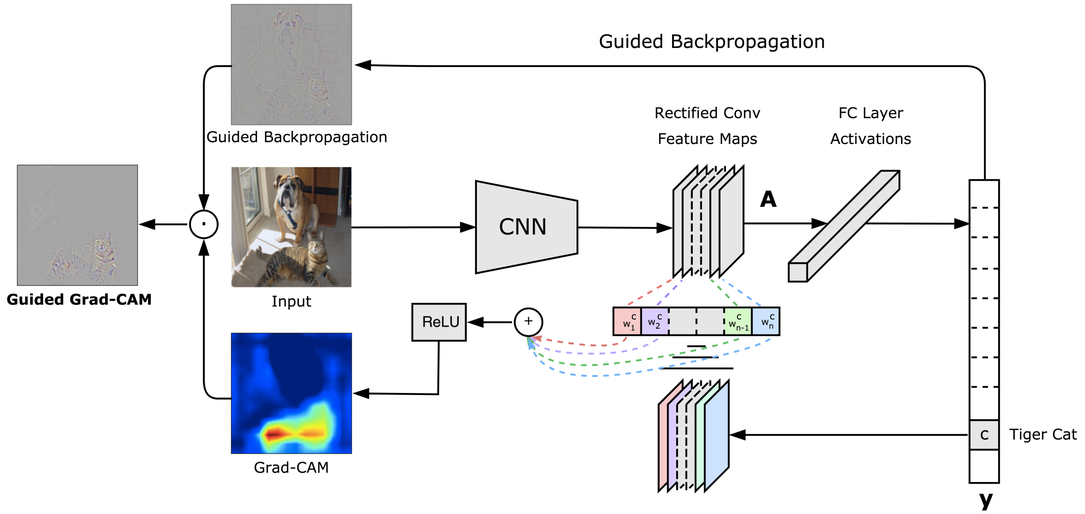
Backdrop-based Methods: + Activations

Class Activation Map (CAM) [Zhou et al., 2016]



* Requires specific architecture: GAP + FC after convs

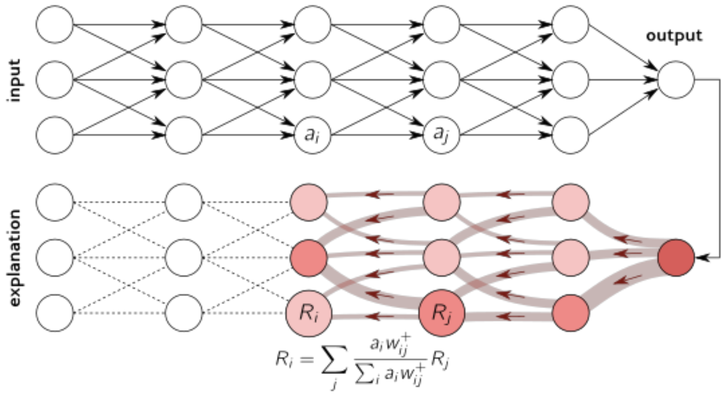
Grad-CAM [Selvaraju et al., 2017]



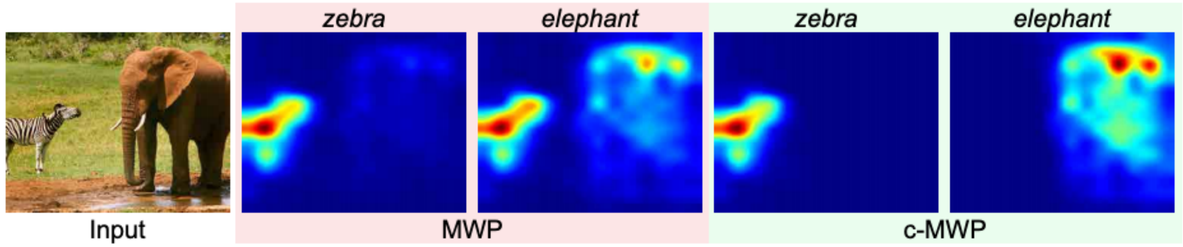
Generalization of CAM for any architecture

Backdrop-based Methods: Conservation Principle (a.k.a. sum to 1)

Layer-wise Relevance Propagation [Bach et al., 2015]

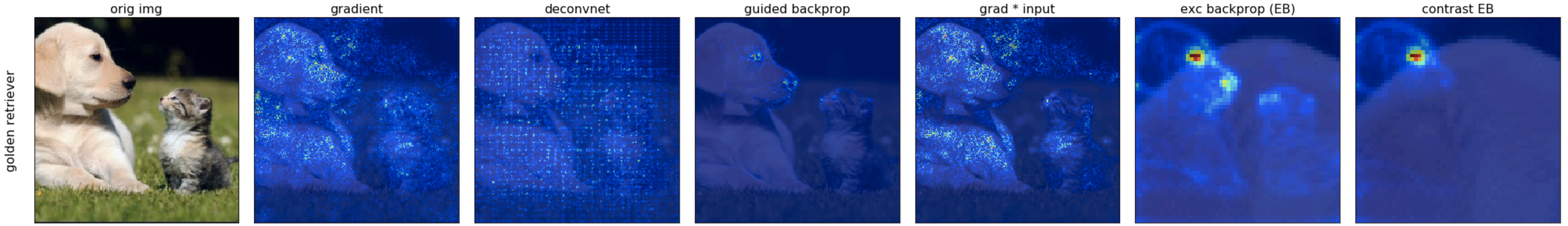


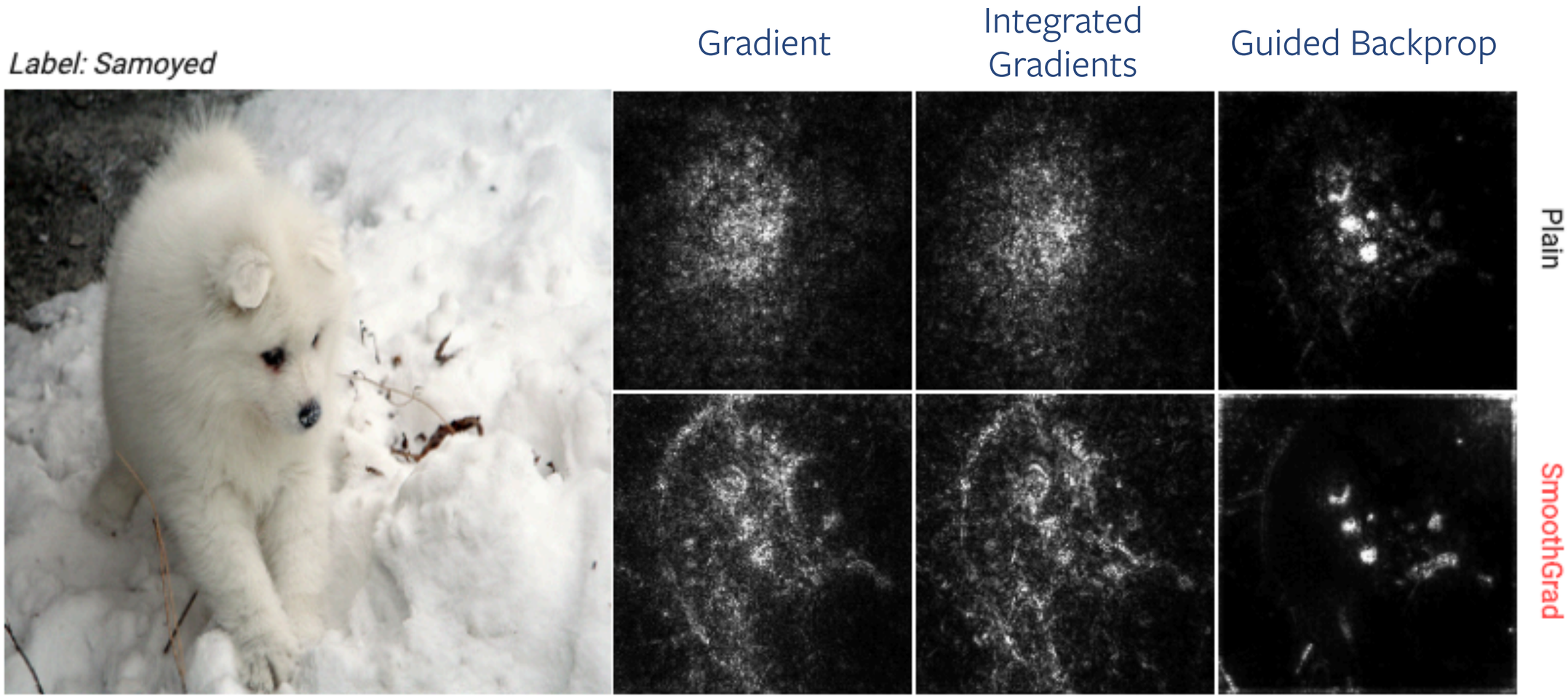
Excitation Backprop [Zhang et al., 2016]



* Includes “contrastive” variant

Both require custom backward functions for most kinds of layers


orig img
gradient
deconvnet
guided backprop
grad * input
exc backprop (EB)
contrast EB
golden retriever

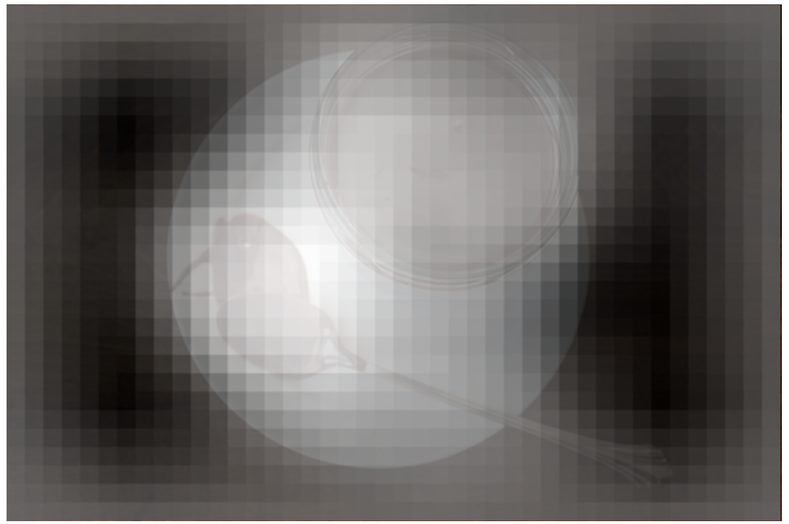

Label: Samoyed
Gradient
Integrated Gradients
Guided Backprop
Plain
SmoothGrad

Perturbation

Change the input and observe the effect on the output

Input

Occlusion



RISE

Clear meaning, but can only test a small number of occlusion patterns

[Zeiler and Fergus, ECCV 2014; Petsiuk et al., BMVC 2018]

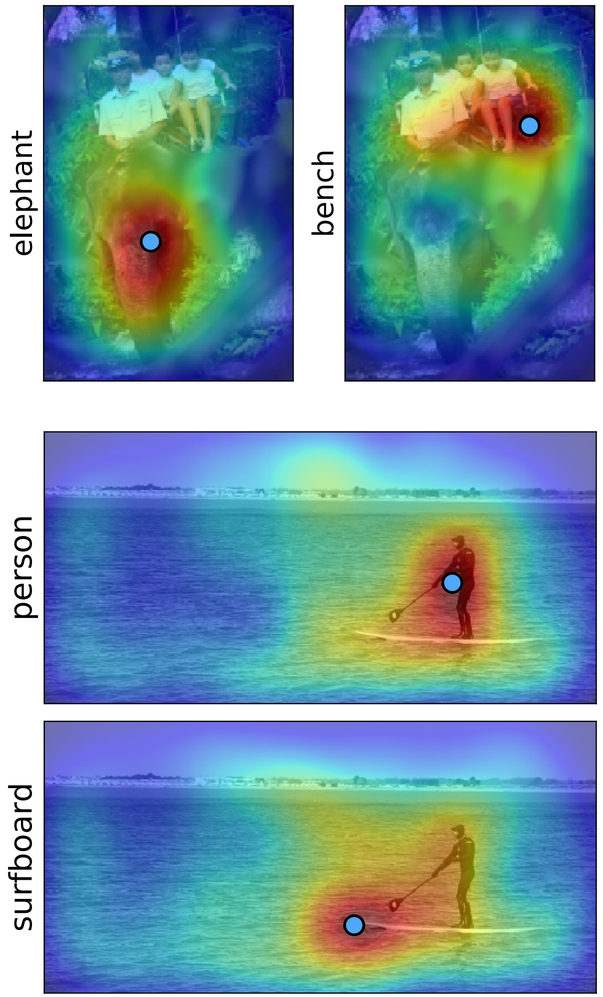
What are desiderata for attribution?

How would one evaluate an attribution method?

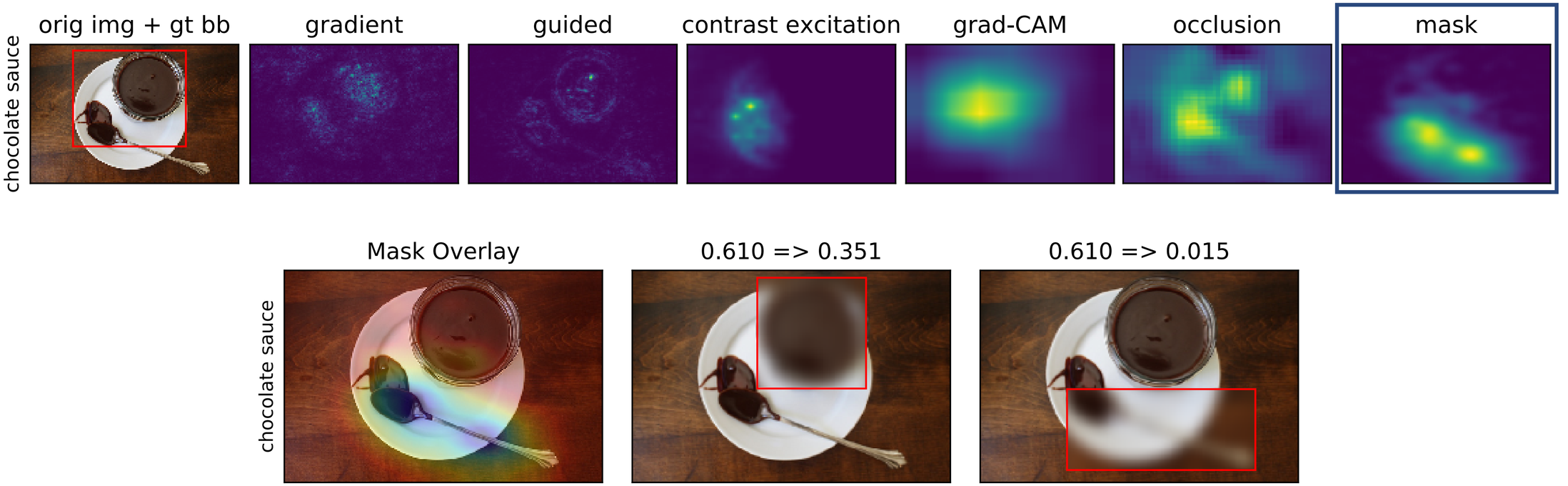
0. Measure Performance on Weak Localization

a. ImageNet Bounding Box Localization

b. Pointing Game [Zhang et al., ECCV 2016]



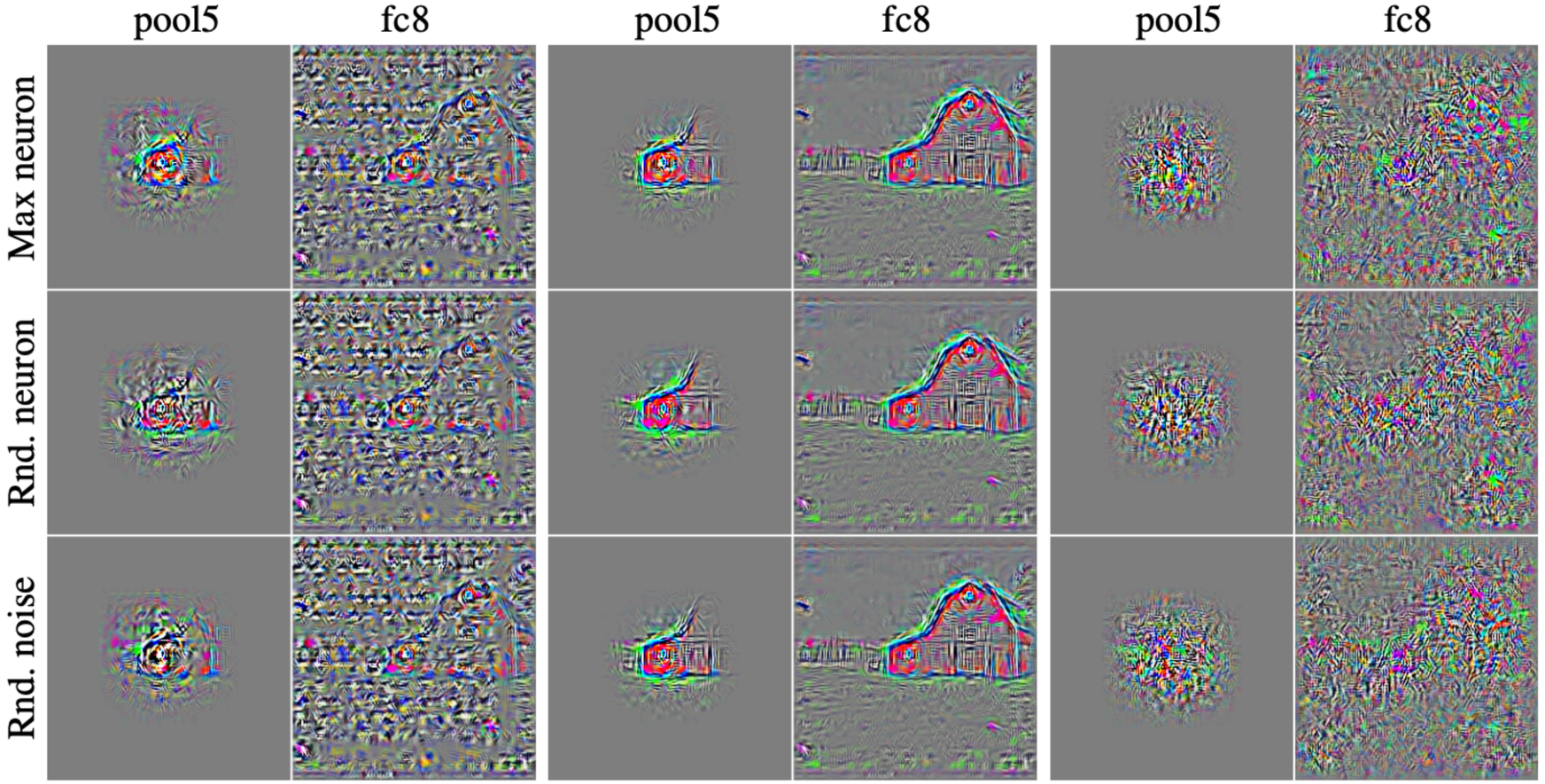
0. Is “Contestable”



An explanation should be **falsifiable**.

[Fong & Vedaldi, ICCV 2017]

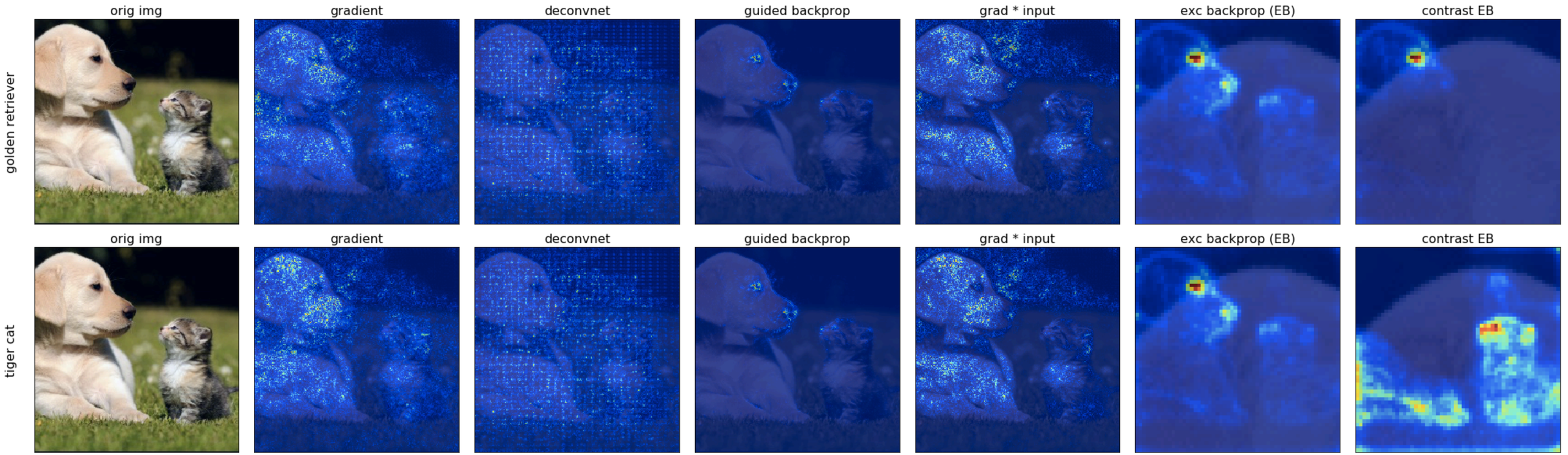
1. Selective to Neuron



DeConvNet Guided Backprop Gradient

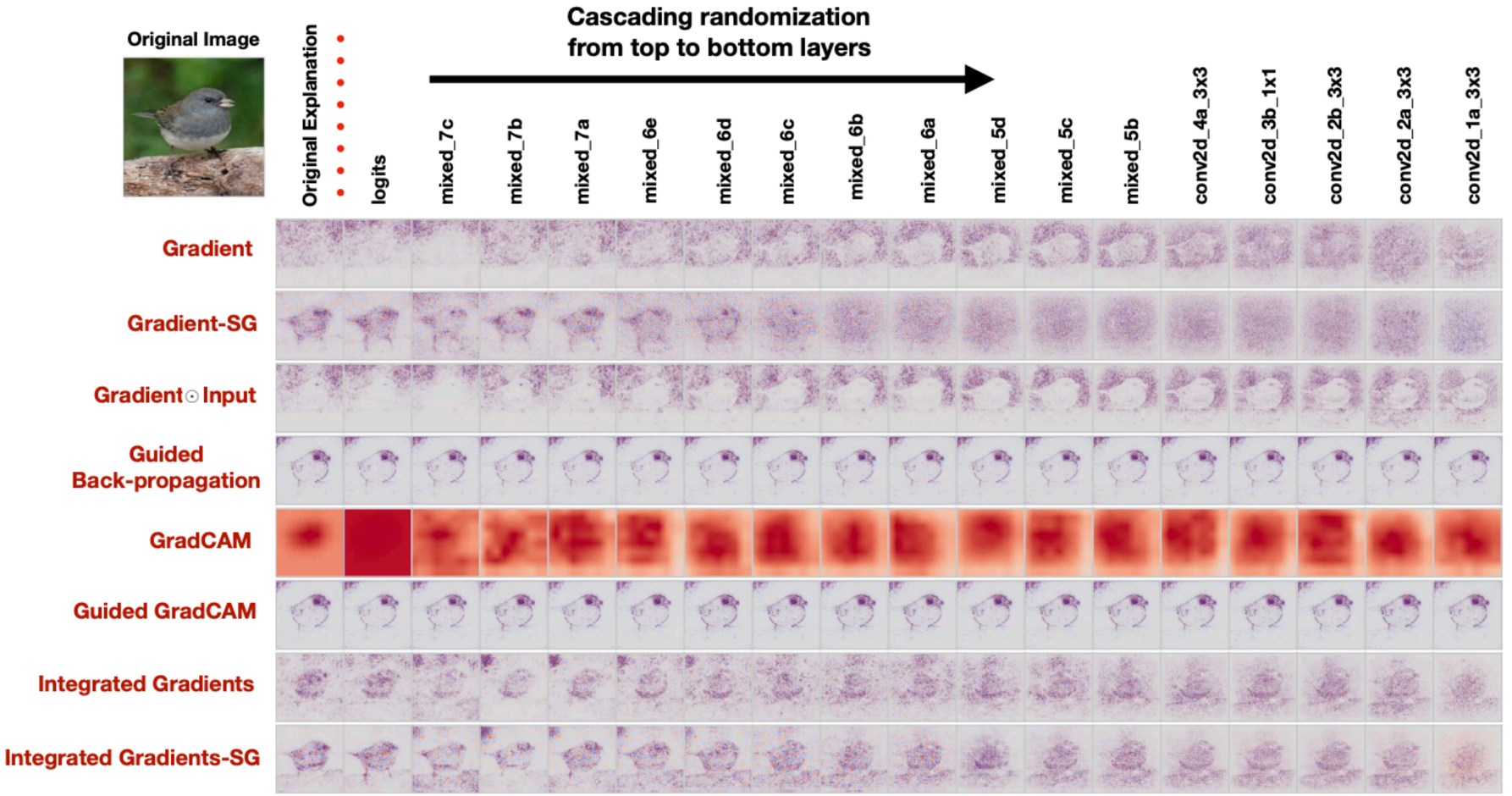
[Mahendran & Vedaldi, ECCV 2016]

1. Selective to Neuron



[Mahendran & Vedaldi, ECCV 2016]

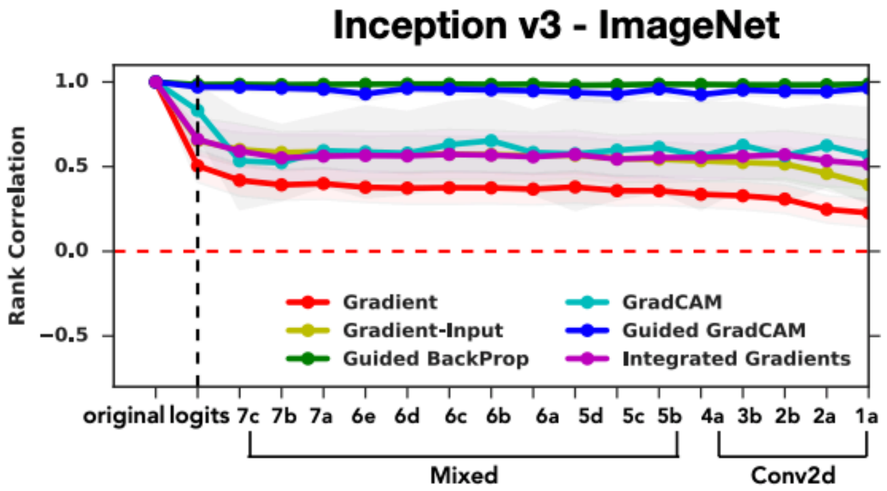
2. Sensitive to Model Parameters

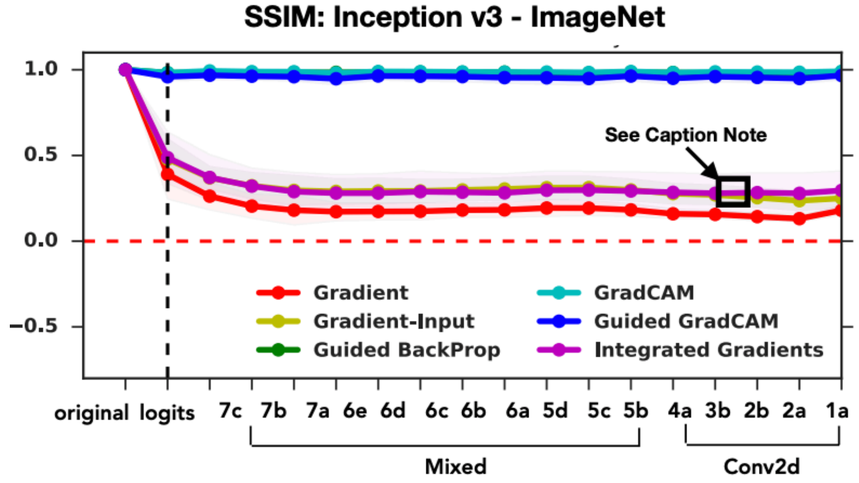


[Adebayo et al., NeurIPS 2018]

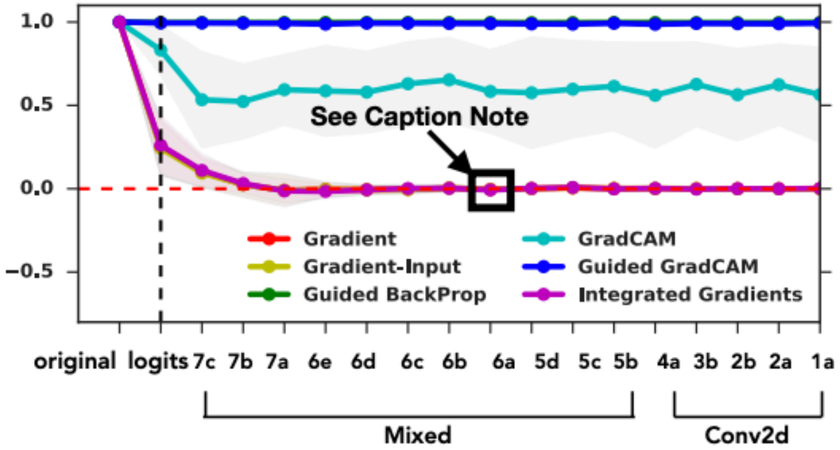
2. Sensitive to Model Parameters

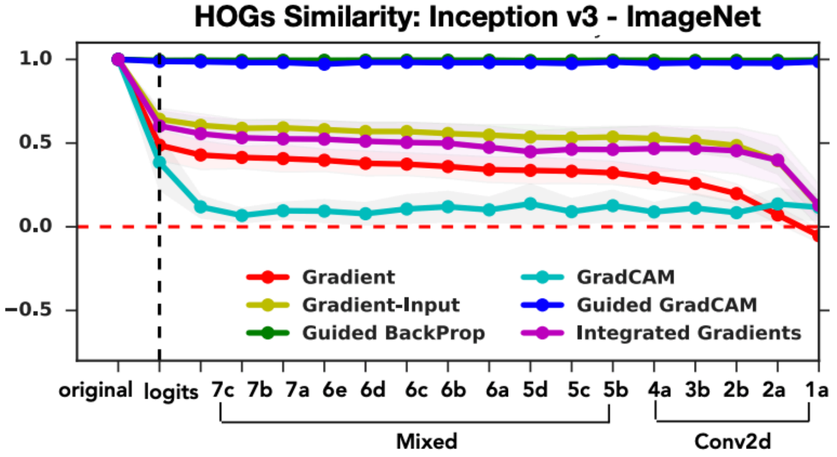
Rank Correlation ABS





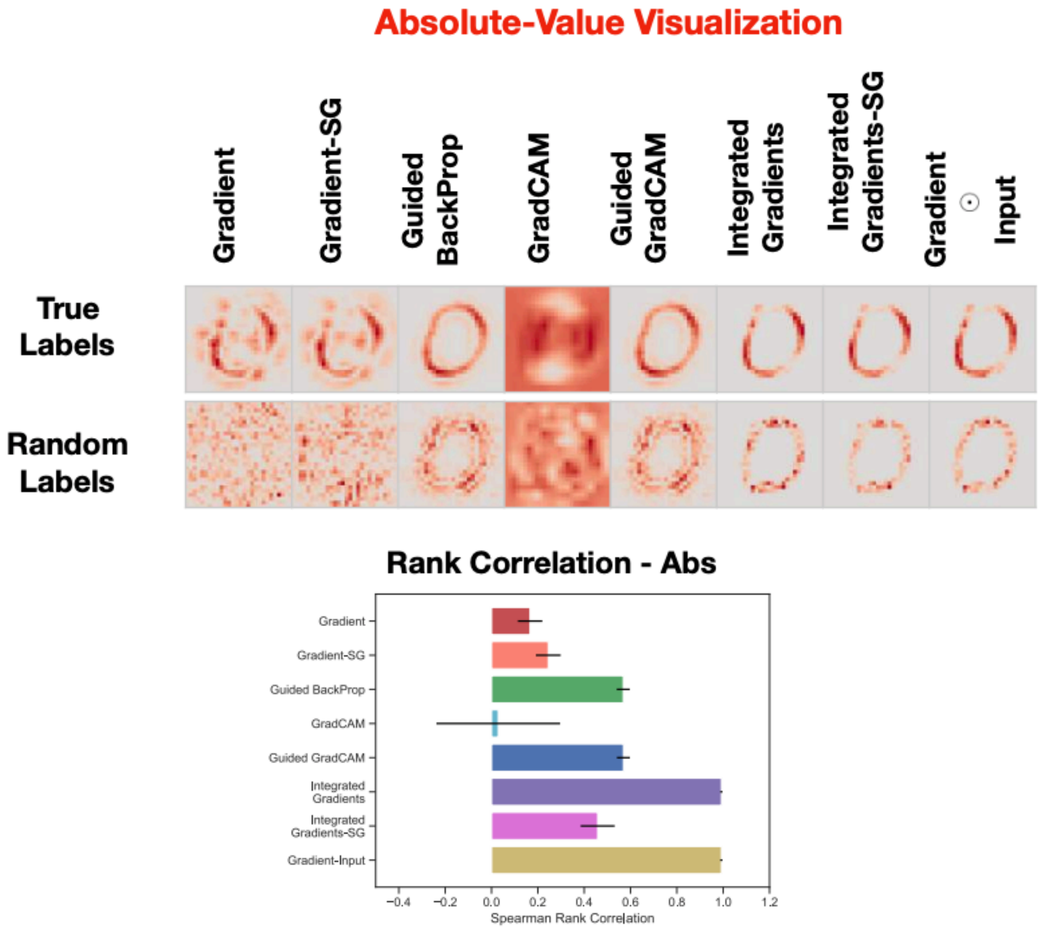
Rank Correlation No ABS

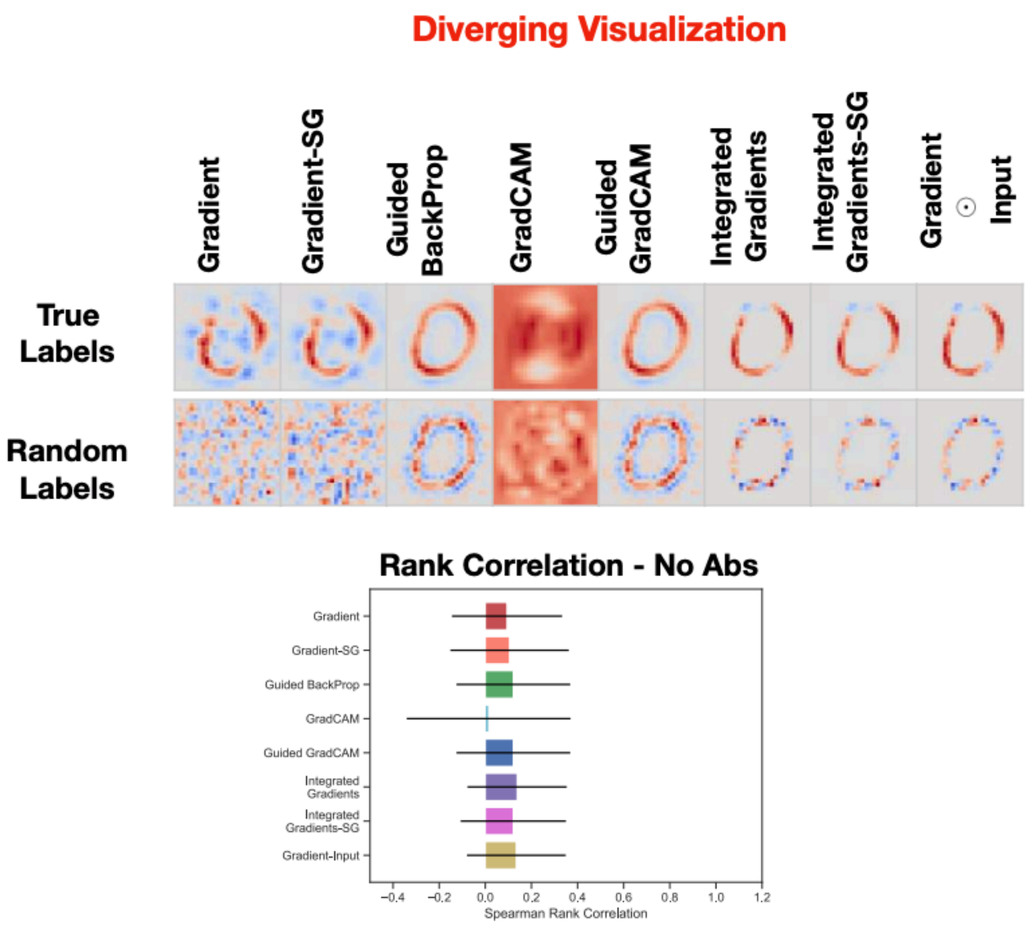




[Adebayo et al., NeurIPS 2018]

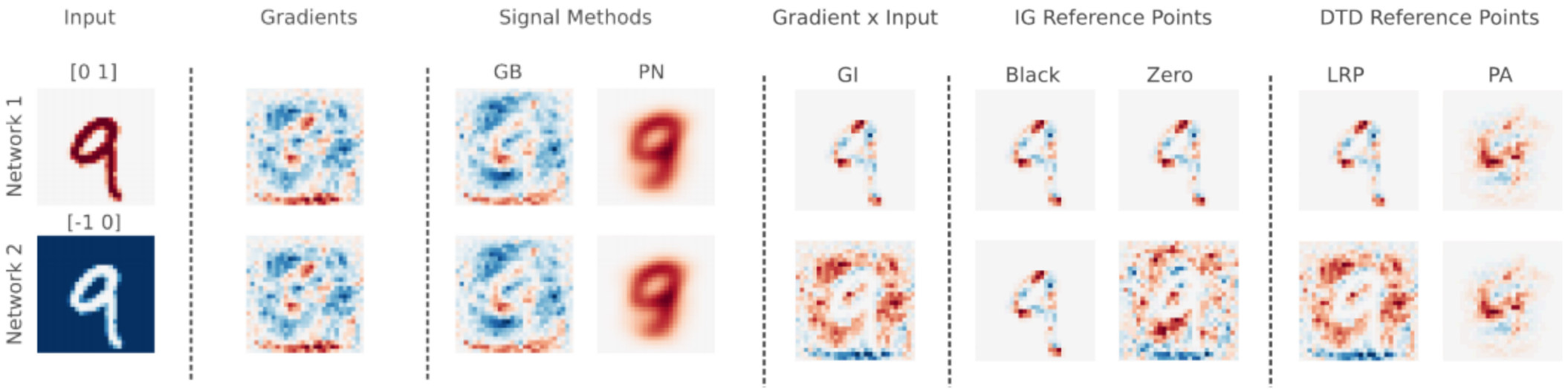
3. Sensitive to Data Labels





[Adebayo et al., NeurIPS 2018]

3. Shift Invariant



[Kindermans et al., arXiv 2017; Zhang, ICML 2019]

4. Perturbation Analysis

Deletion Game

a. Successively extract salient patches from heatmap and “delete” them

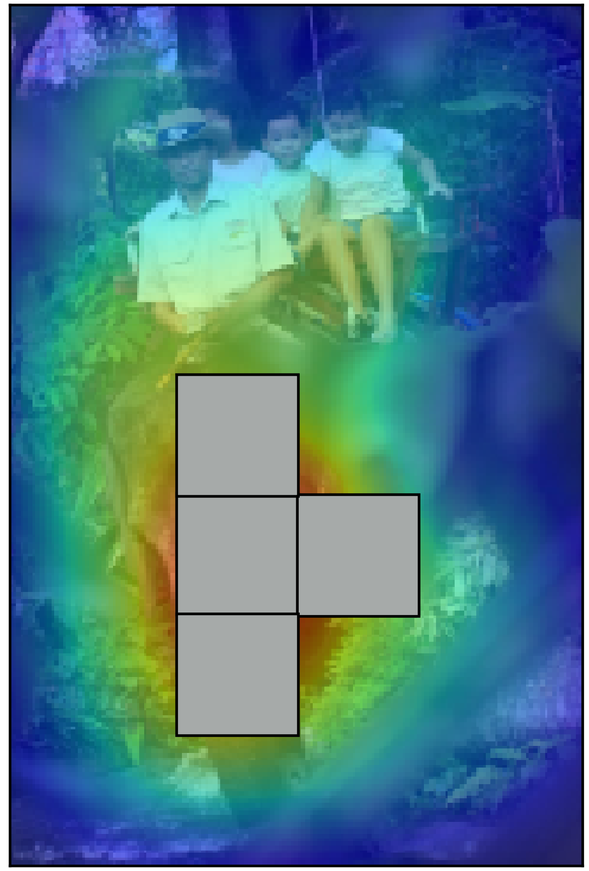
b. Plot curve and report AUC

Problems?

* Choice in patch size

* Evaluating outside of training domain

ROAR (**R**em**o**ve **A**nd **R**etrain) [Hooker et al., 2018]
* Retrain classifiers with X% of features (i.e., pixels) removed



Best practices

* Assume the model has failure modes and seek to explain them with attribution methods

* If using backdrop-based methods, consider contrastive methods:
Contrastive Excitation Backprop
Competitive Gradient * Input [Gupta and Arora, 2019]

* Don’t use Guided Backprop to “improve visual quality”

What should interpretability research focus on from hence forth?

See white paper.

Future directions

1. Common benchmarks and standards for evaluating desiderata for interpretability properly
2. Tools for practitioners
3. Under-explored research areas

TorchRay

github.com/facebookresearch/torchray

PyTorch