SmoothGrad: removing noise by adding noise

Paper Authors: Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viegas, Martin Wattenberg

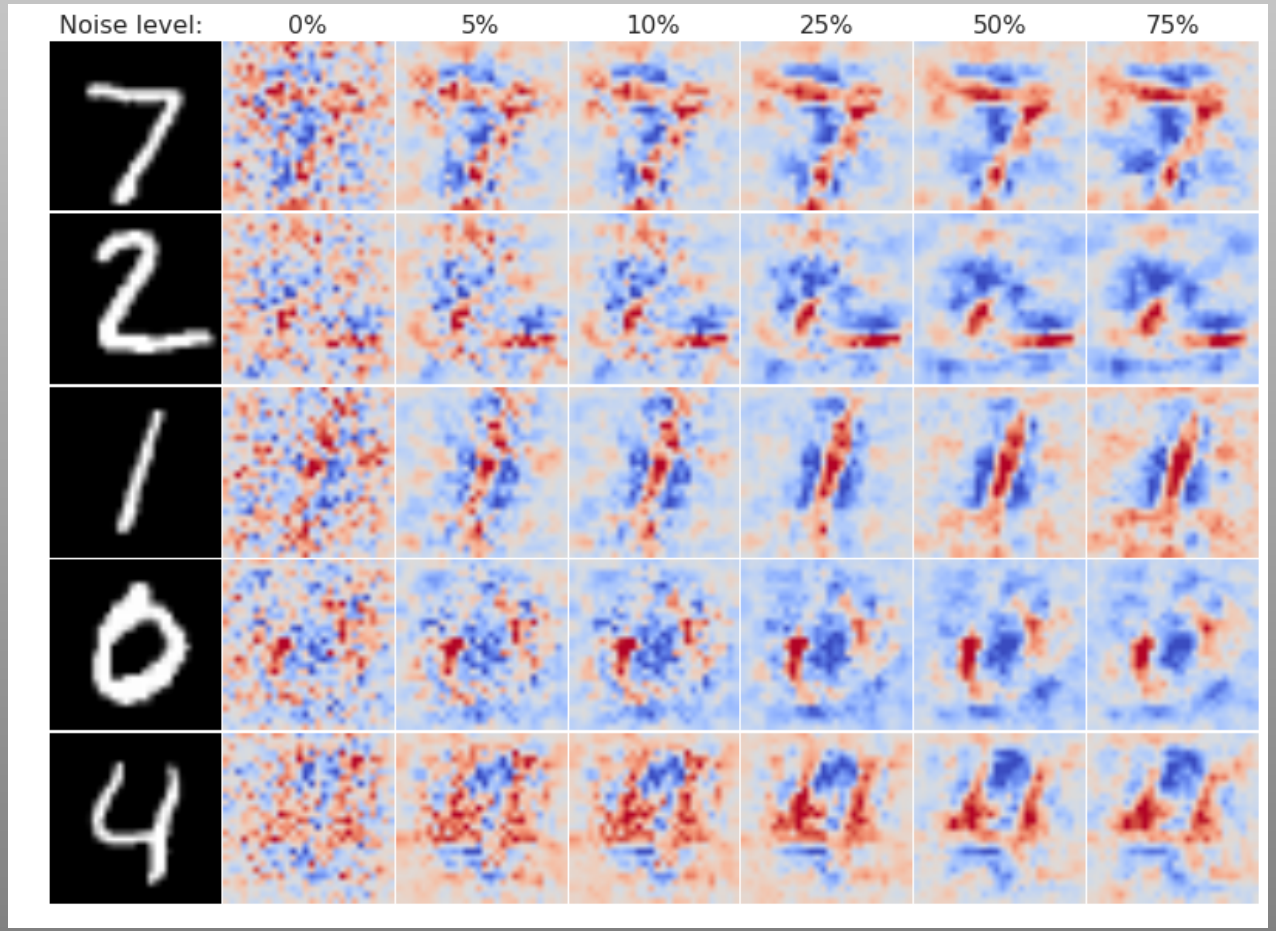


Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Motivation

- One kind of explanation: Identify pixels that lead to the DNN decision.
- Starting point
  - Gradient of the class score or logit w.r.t. input.
  - Sort of a *sensitivity map*
- Two contributions
  - SMOOTHGRAD
    - visually sharper sensitivity maps
  - Lessons in the visualization of these maps
- Artifacts
  - Code
  - Website

Gradients as sensitivity maps

- A DNN that classifies
  - an input image x
  - into one class c
  - from a set C of possible classes
- DNN computes a class activation function S_c for each class $c \in C$
- The final classification $class(x)$ determined by the highest score.
- That is,

$$class(x) = argmax_{c \in C} S_c(x)$$

Gradients as sensitivity maps - II

- If class activation functions S_c are piecewise differentiable,
- for any image x,
- construct a *sensitivity map* $M_c(x)$ by differentiating M_c w.r.t. the input x.

$$M_c(x) = \partial Sc(x)/\partial x$$

  - M_c describes how a change in a pixel of x impacts its label as class c
  - Mathematically rigorous method of allocating importance to pixels
- Sensitivity maps of raw gradients are visually noisy
- Poor correlation with human expectation

Erhan D, Bengio Y, Courville A, Vincent P. Visualizing higher-layer features of a deep network. University of Montreal. 2009 Jun 9;1341(3):1.

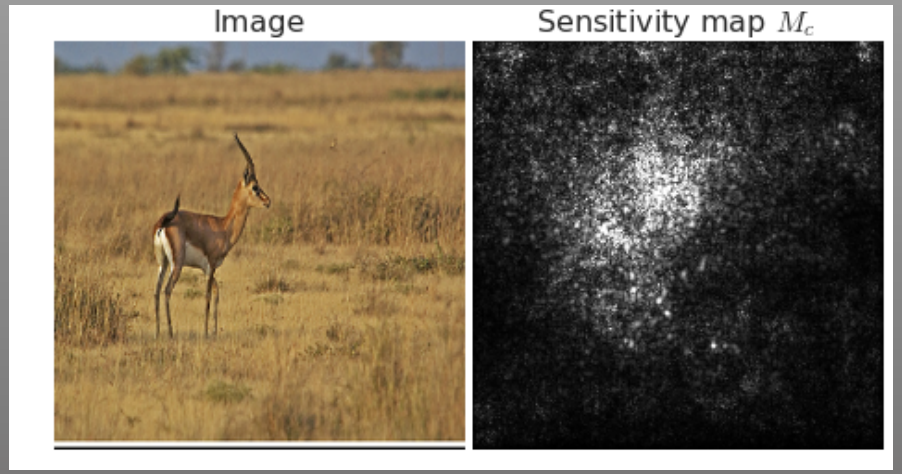


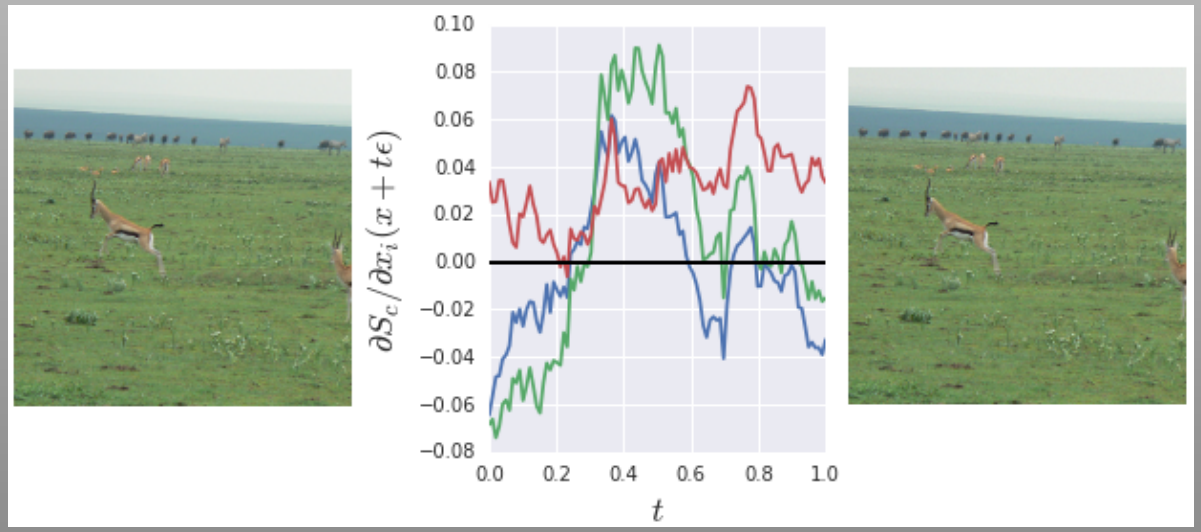
Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Enhanced sensitivity maps

- Hypotheses for noisy raw gradients:
  - Honest to what the network has learned
  - Not an effective proxy for feature importance
- Attempts at many sensitivity maps:
  - Features may “saturate”; strong effect globally, but with a small derivative locally.
    - *Layerwise Relevance Propagation* (Bach et al., 2015)
    - *DeepLift* (Shrikumar et al., 2017)
    - *Integrated Gradients* (Sundararajan et al., 2017)
  - Extend backpropagation and emphasize positive contributions
  - Modify gradients of ReLU discarding negative values during backpropagation
    - *Deconvolution* (Zeiler & Fergus, 2014)
    - *Guided Backpropagation* (Springenberg et al., 2014)

Smoothing noisy gradients

- Potential explanation
  - the derivative of the class activation function S_c may fluctuate sharply
  - essentially meaningless local variations in partial derivatives.
  - ReLU activations
    - S_c not even continuously differentiable
- Gradient of S_c at any given point $\frac{\partial S_c}{\partial x_i}(t)$ less meaningful than a local average of gradient values
  - Smoothen ∂S_c with a Gaussian kernel
  - Computing an average intractable
    - High-dimensional inputs



Plot of the values of $\frac{\partial S_c}{\partial x_i}(t)$ as fraction of the maximum $\max_i \frac{\partial S_c}{\partial x_i}(t)$ for a segment $x + t\varepsilon$ in the space of images.

Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Smoothening noisy gradients - II

- Stochastic approximation SMOOTHGRAD:
  - Take random samples in the neighborhood of an input x,
  - Average the resulting sensitivity maps.
- Mathematically,

$$\hat{M}_c(x) = \frac{1}{n}\sum_{1}^{n} M_c(x + \mathcal{N}(0, \sigma^2))$$

- Here,
  - n is the number of samples
  - $\mathcal{N}(0, \sigma^2)$ represents Gaussian noise with standard deviation σ.

Experiments

- Two image classification models:
  - Inception v3 model (Szegedy et al., 2016)
  - a convolutional MNIST model
- Smoothed gradient, M_c, visually more coherent
- Sign of gradients in heat map visualizations:
  - MNIST: positive gradients indicate support for the class
  - ImageNet: absolute value leads to clearer pictures
    - direction is context dependent
    - image recognition invariant under
      - color changes (?)
      - illumination changes
- Outlier removal in heat maps:
  - Bounding values to 99th percentile is visually coherent
- Multiplying maps with input images:
  - May borrow clarity from the input.
  - In a linear system $y = W x$, product makes sense

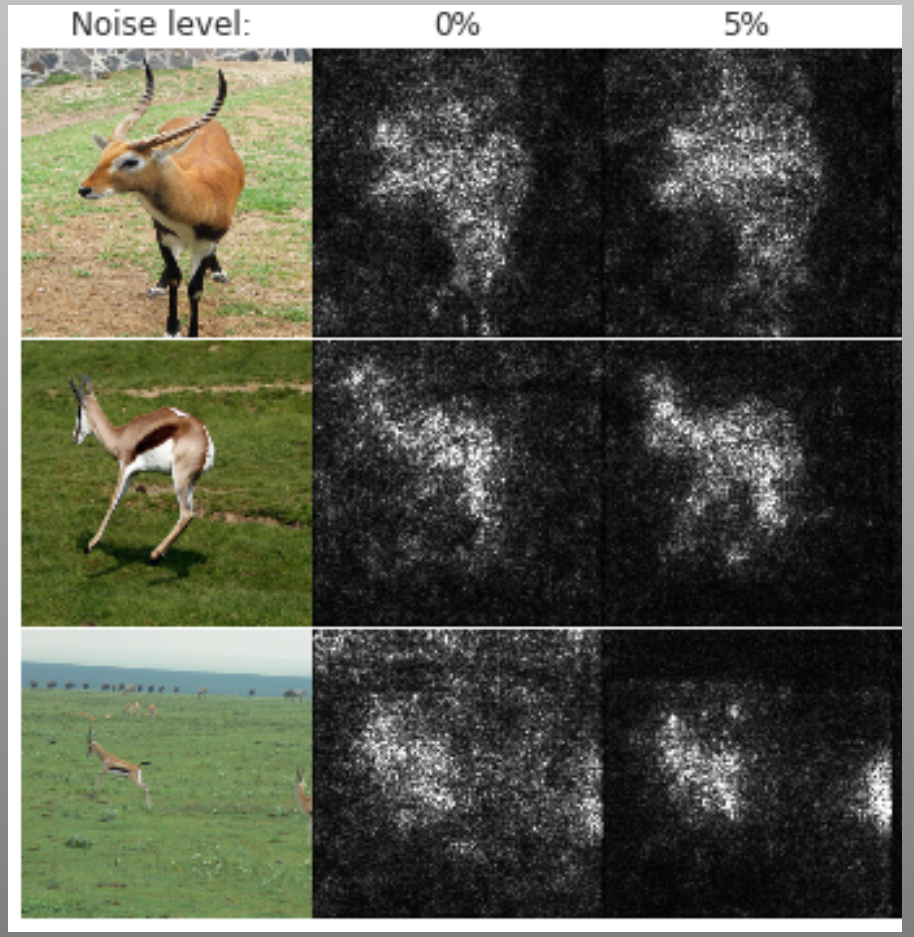


Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Impact of Noise on Attribution

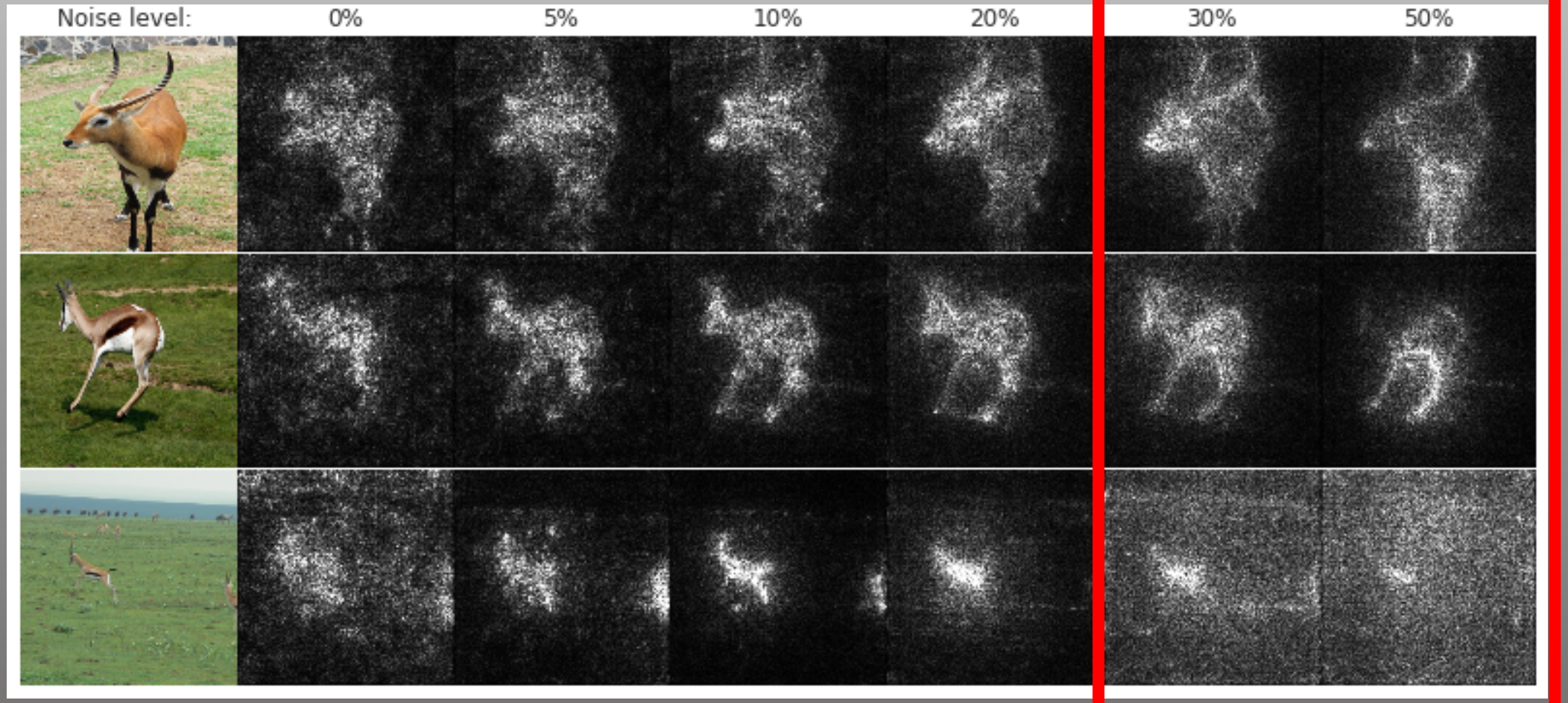


Image reproduced under fair use from
https://arxiv.org/pdf/1706.03825.pdf

Noise and sample size of SMOOTHGRAD

- **Noise, σ**
  - 10%-20% noise balances sharpness and structure of the original image.
  - Ideal noise level depends on the input.

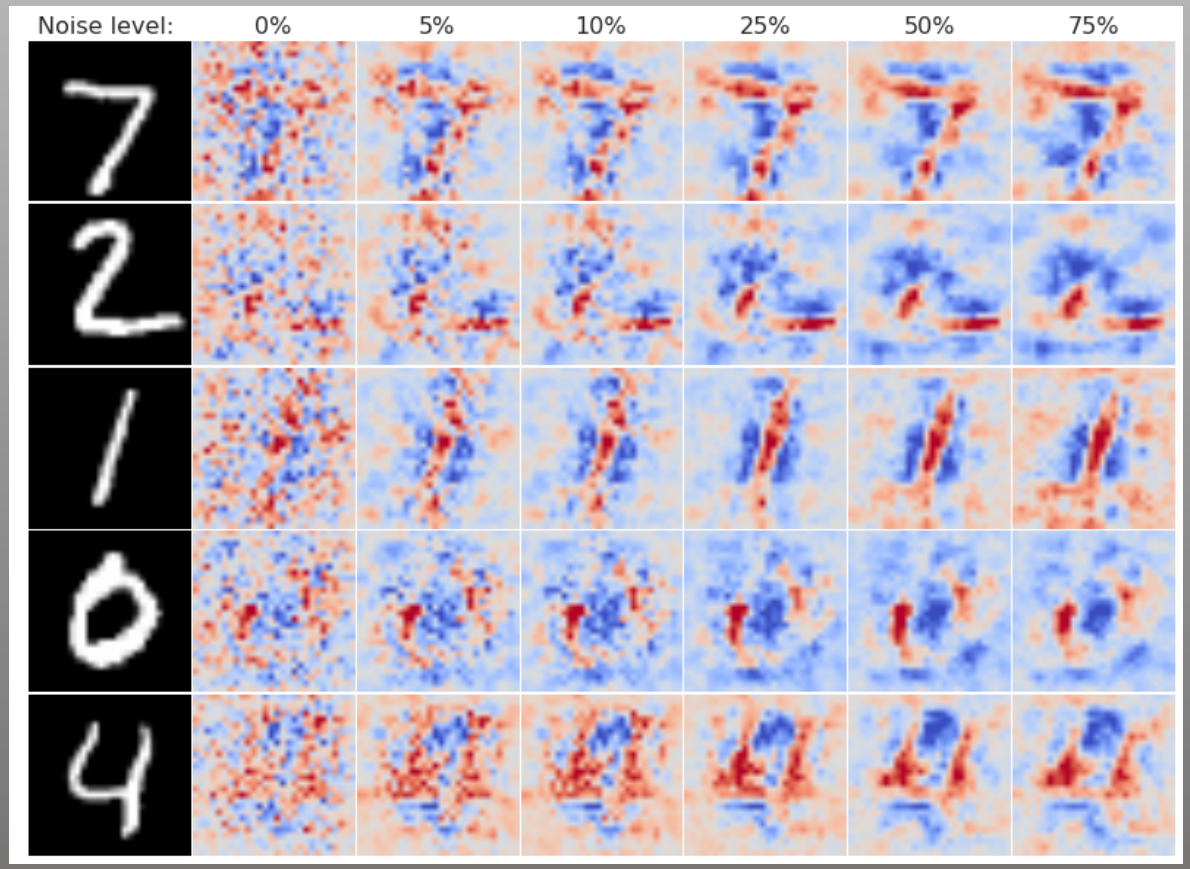


Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Noise and sample size of SMOOTHGRAD - II

- **Sample size, n**
  - estimated gradient is smoother as sample size, n, grows in size.
  - diminishing return for $n > 50$

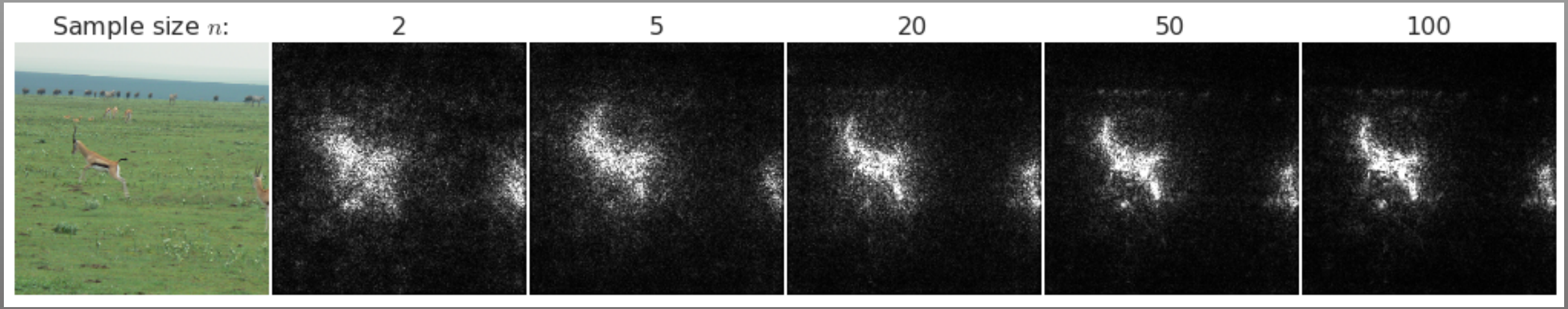


Image reproduced under fair use from
https://arxiv.org/pdf/1706.03825.pdf

Evaluation: visual coherence

- Compared with three gradient-based methods:
  - *Integrated Gradients* (Sundararajan et al., 2017),
  - *Guided BackProp* (Springenberg et al., 2014)
  - vanilla gradient.
- Visual self-inspection of 200 images
- Guided Backprop sharper
  - But prone to failure

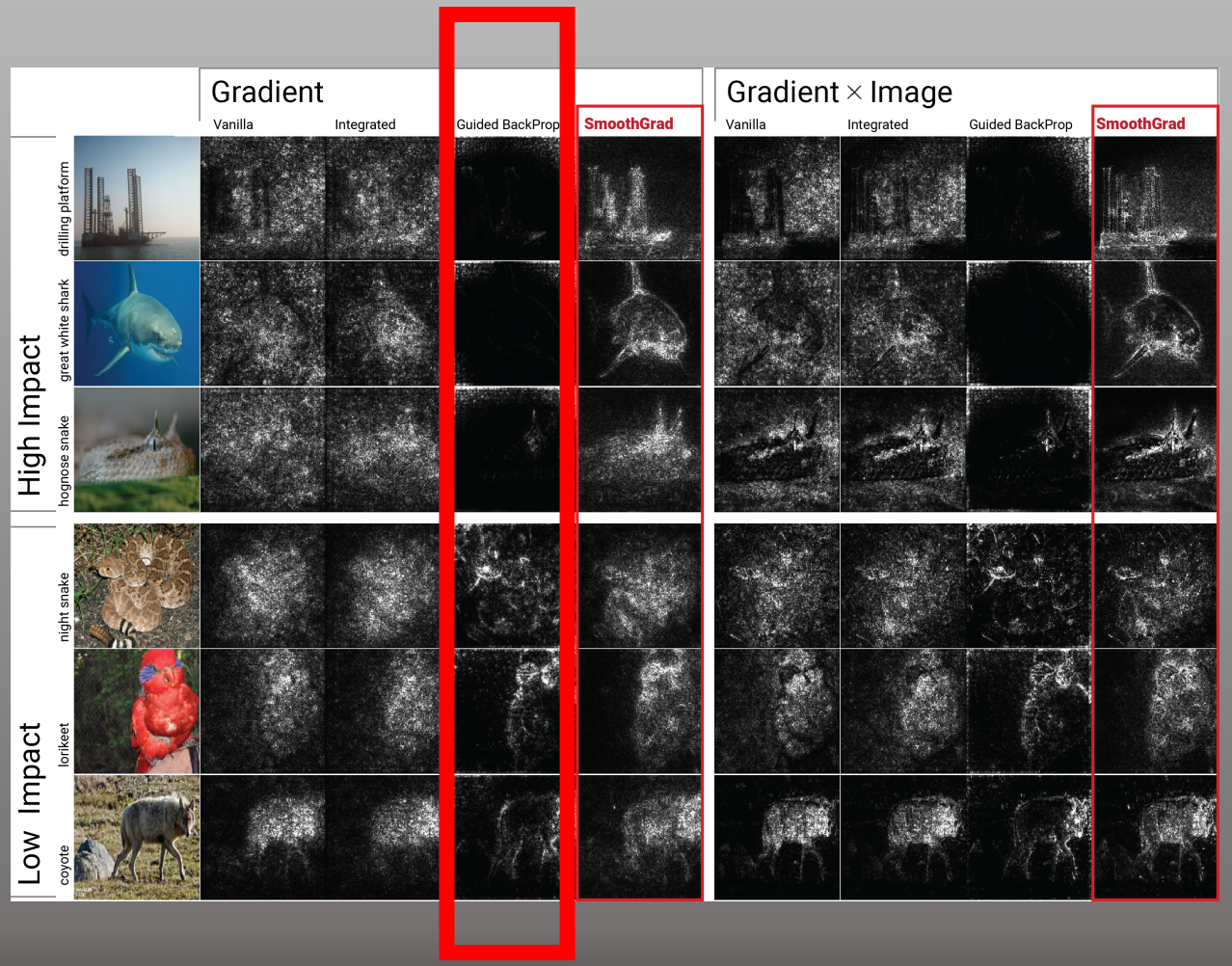


Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Evaluation: discriminativity

- Choose images with at least two objects of different classes.
- Compute the sensitivity maps M1(x) and M2 (x) for both classes
- Scale both to [0, 1], and calculate the difference M1 (x) − M2 (x).
- Plot the values on a diverging color map [−1, 0, 1] → [blue, gray, red].

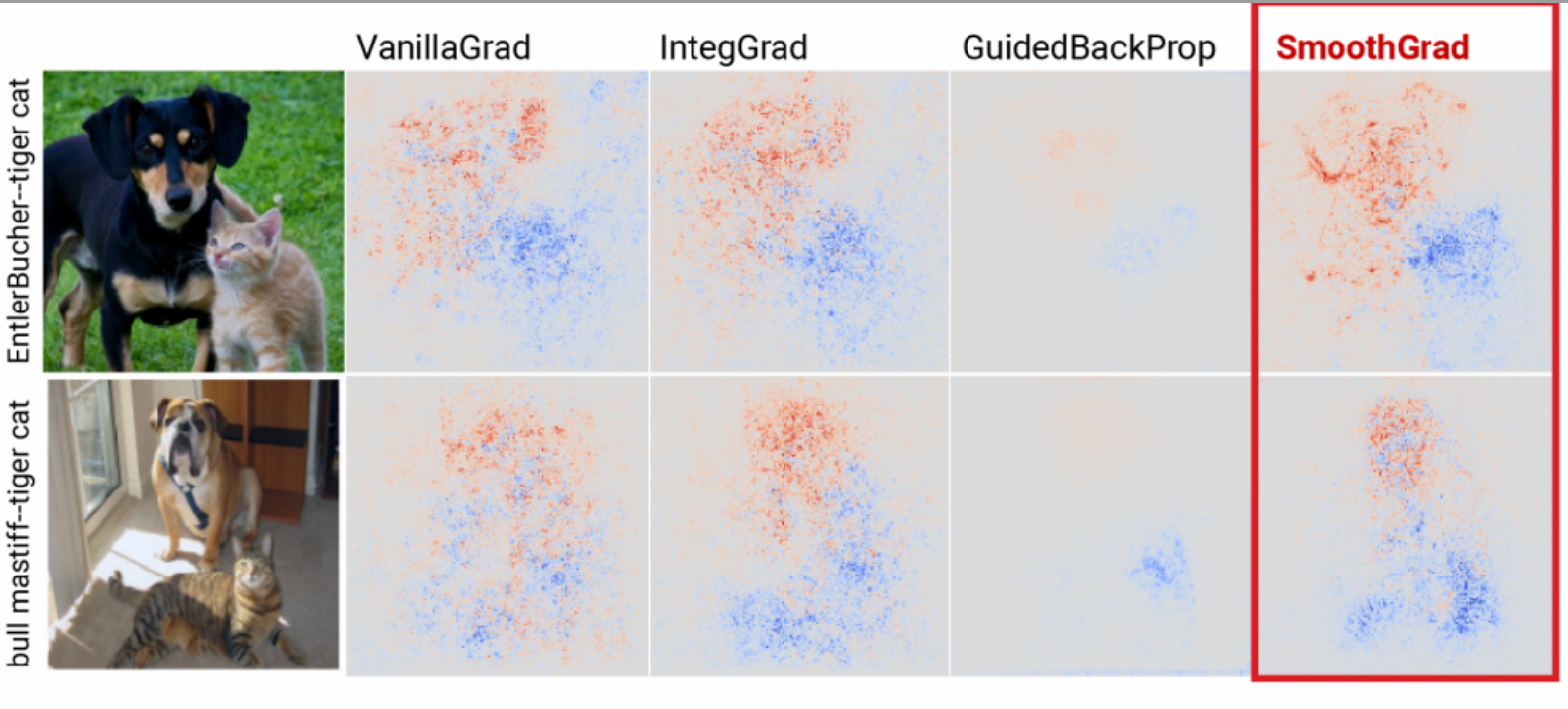


Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

SmoothGRAD + IG, Guided BackProp

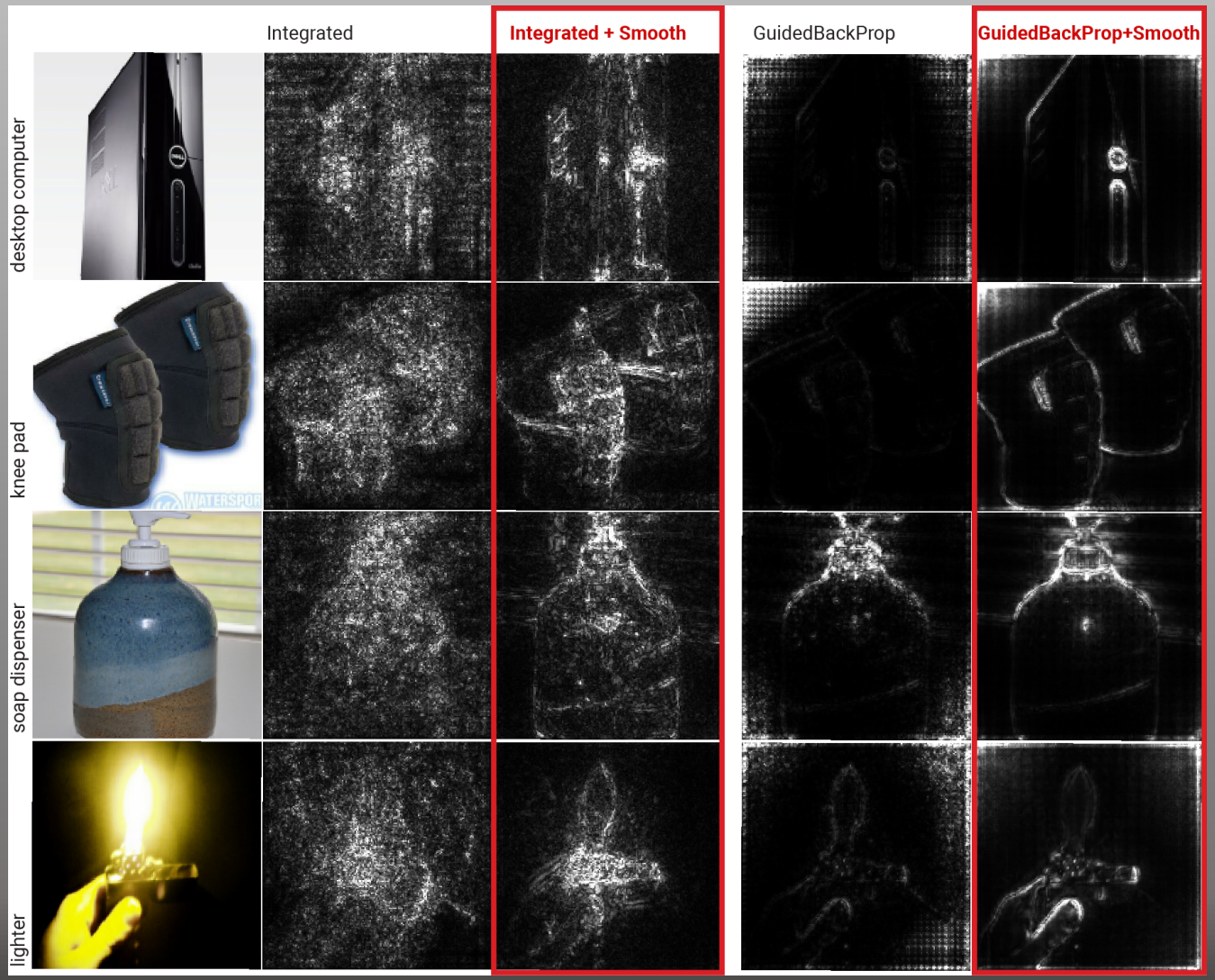


Image reproduced under fair use from https://arxiv.org/pdf/1706.03825.pdf

Conclusions

- Averaging maps of image + perturbations smoothens explanations
- Effect enhanced further by training on data with random noise
- Future Work:
  - Investigate if noisy sensitivity maps arise due to noisy gradients?
    - Theoretical arguments
  - Other explanations for SMOOTHGRAD
    - random noise and its interactions with different textures
  - Direct methods to learn DNNs with smoother class score functions
    - Penalty on large partial derivatives
    - Explicit penalty for changes in derivatives of the class score w.r.t. neighboring pixels
  - Understand the geometry of the class score function
    - Explain why smoothing is better with large areas of near-constant pixel values?
  - Better metrics and data sets