SmoothGrad: Removing Noise by Adding Noise

Presentation By:
Eric Watson & William Sawran

Paper Details

Authors:

- Daniel Smilkov
- Nikhil Thorat
- Been Kim
- Fernanda Viegas
- Martin Wattenberg

Published: ICML, 2017 Workshop on Visualization for Deep Learning

Citations: 553

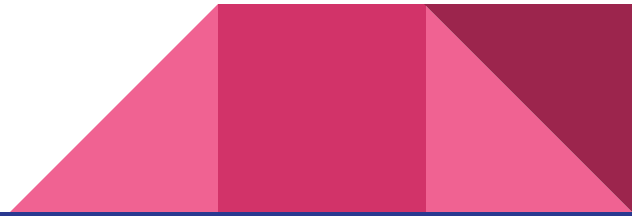
Paper: https://arxiv.org/pdf/1706.03825.pdf

Website: https://pair-code.github.io/saliency/

Code: https://github.com/pair-code/saliency

Overview

- Definition of Sensitivity Maps
- Previous Work
- SmoothGrad Proposal
- Experiments
- Conclusion
- For and Against Paper



Definition of Sensitivity Maps

- A sensitivity map is used to visualize the important features of a prediction from a machine learning model
- Model Prediction:

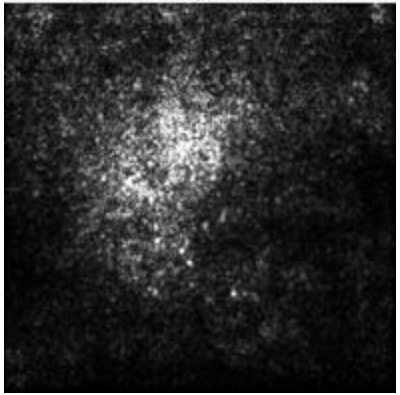
$$class(x) = \text{argmax}_{c \in C} S_c(x)$$

- Sensitivity Map:

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

Image

Sensitivity map M_c



Previous Work

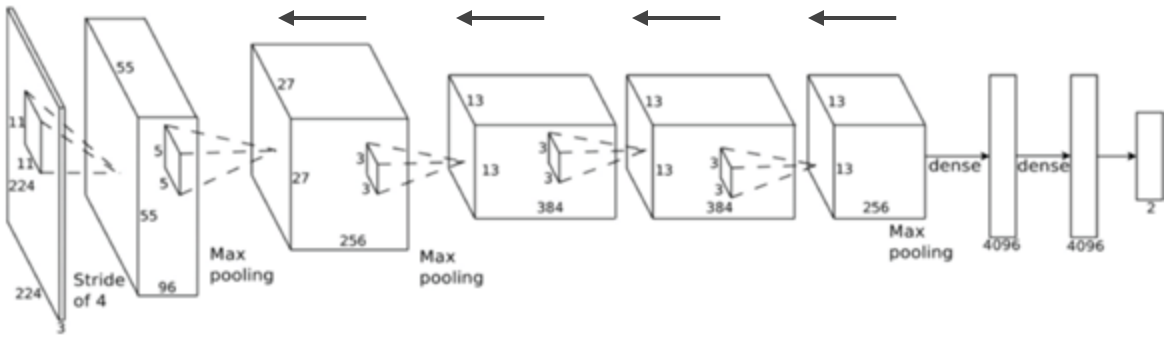
- Sensitivity maps can visually highlight the pixels of informative features
- Previous proposals have been made to enhance sensitivity maps
- Layerwise Relevance Propagation, DeepLift, and Integrated Gradients
  - Referred to as saliency maps or pixel attribution maps
  - Estimate the global importance of each pixel rather than local sensitivity

Integrated Gradients - Linearly Interpolated Input Images

Previous Work

- Deconvolution and Guided Backpropagation
  - Modifies the backpropagation algorithm

Zero out negative gradients during backpropagation



SmoothGrad Proposal

- Why do sensitivity maps highlight noise?
- Noise may be due to local variations in partial derivatives of the S_c function
- In order to present the possible explanation, the authors derived an example

SmoothGrad Proposal

- Where x is the image, and $x + \epsilon$ is the final image
- Where ϵ is one random gaussian noise sample from: $N(0, 0.01^2)$
- Where the middle graph is the maximum entry in the gradient: $\max_i[\partial S_c/\partial x_i]$ (t) for the input of $x + t\epsilon$ in the space parameterized by $t \in [0, 1]$

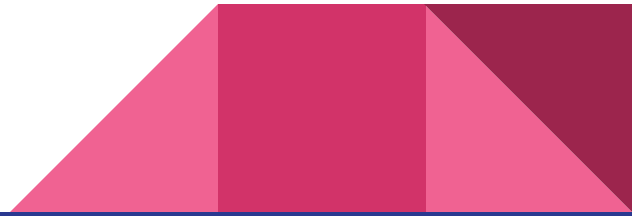
SmoothGrad Proposal

- Fluctuations in the gradient suggest taking a local average of the gradient values
- However, computing a local average in a high-dimensional input space is computationally expensive
- SmoothGrad leverages a stochastic approximation
- The equation derived for SmoothGrad is the following:

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

Experiments - Models

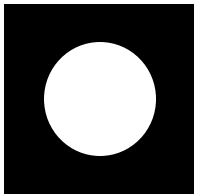
- For experimentation, two neural networks were used for image classification
- First Model:
  - Inception v3
  - Pre-trained on the ILSVRC-2013 dataset
- Second Model:
  - Convolutional MNIST model
  - Based on the TensorFlow tutorial
  - Pre-trained on the MNIST dataset



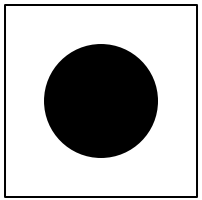
Experiments - Visualization (Value of gradients)

- Heatmaps are typically used for Sensitivity Maps
- However, the channel value for a pixel can impact the output representation
- A choice would be to preserve positive and negative pixel values, or take the absolute value
- MNIST is a case that benefits from the signed values, while ImageNet would benefit from absolute values

Example: Classifying a Ball



Positive Gradient



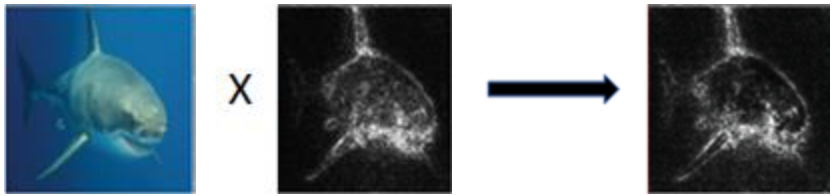
Negative Gradient

Experiments - Visualization (Capping Values)

- Outlier pixels with above-average gradients may throw off the color scales of sensitivity maps
- Capping gradients to a relatively high value produces more visually-coherent maps
- 99th percentile is determined to be sufficient for this

Experiments - Visualization (Multiplying with Input)

- (actual pixel values) x (gradient-based values) = simpler & sharper sensitivity map

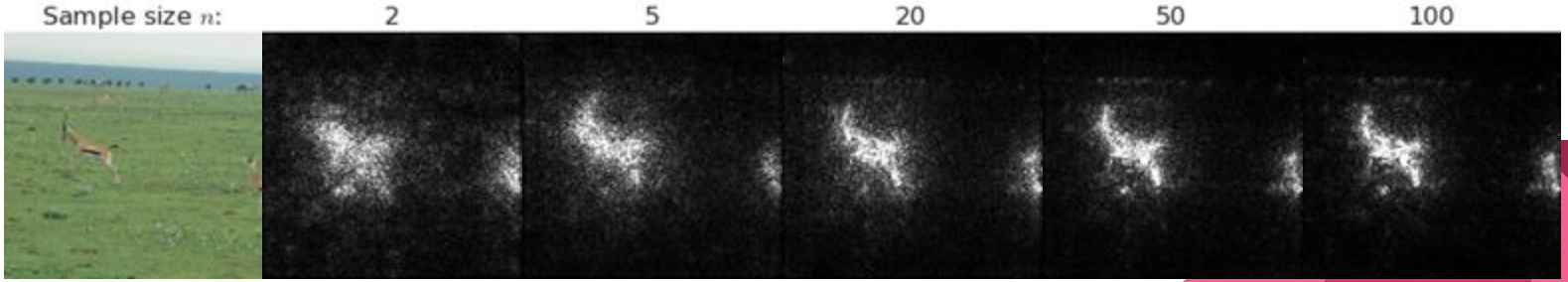


Original Image Gradient Map Final Map

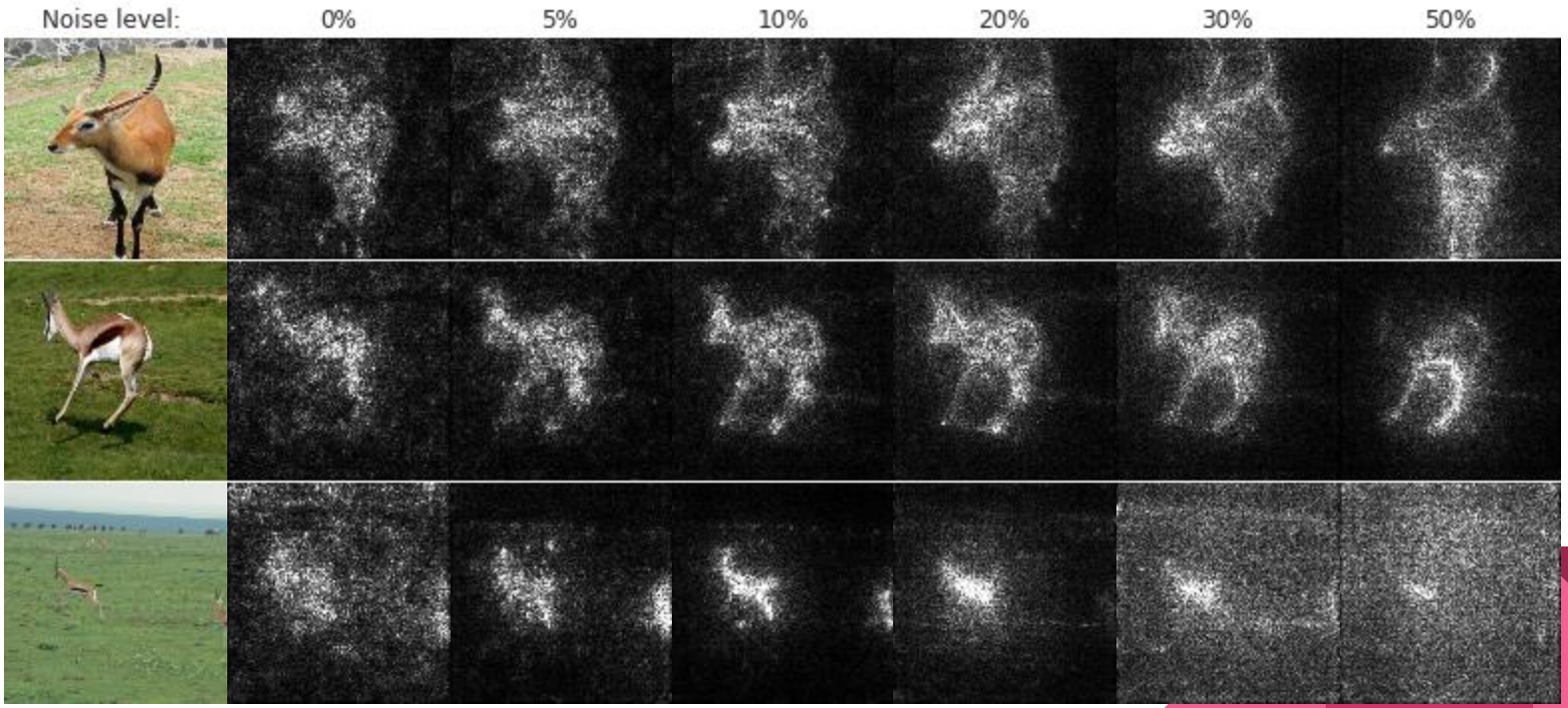
- Like absolute value operation, multiplication can impact output representation
- Hence, the choice of applying multiplication is image-dependent

Experiments - Parameters

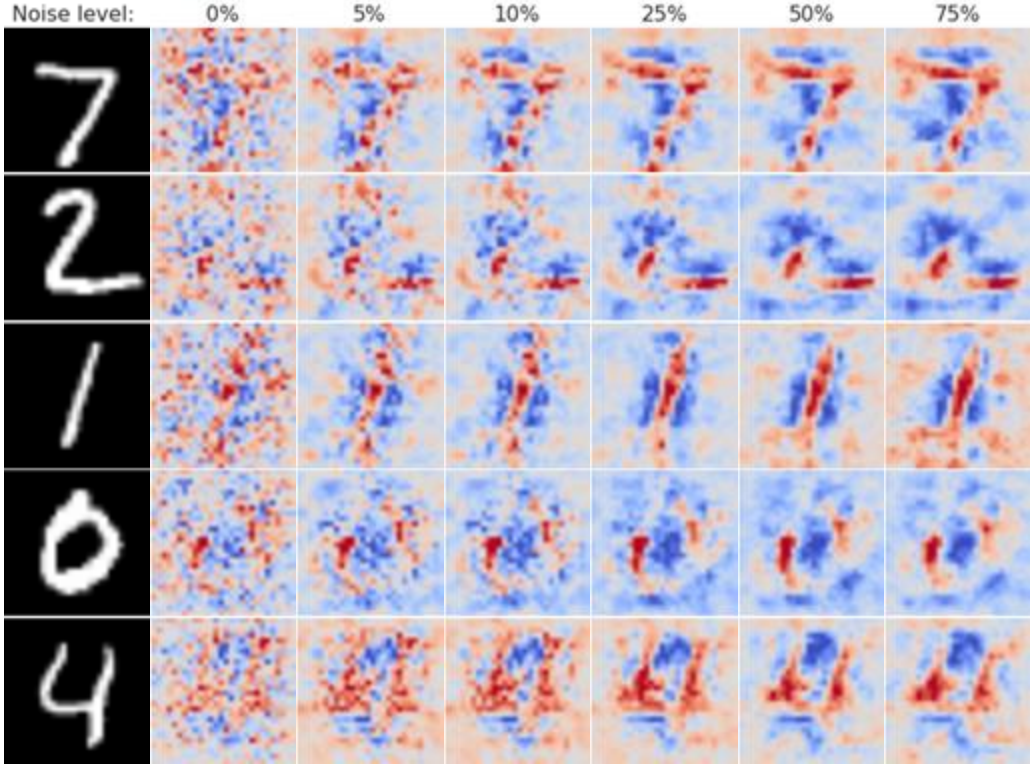
- SmoothGrad uses the two hyper-parameters of σ and n
  - σ controls the noise level of the perturbations
  - n controls the number of samples to average over
- A noise level of (10 - 20)% balances sharpness and structure of the image
- A sample size of 50 provides a smooth gradient, while values above have diminishing return



Experiments - Parameters



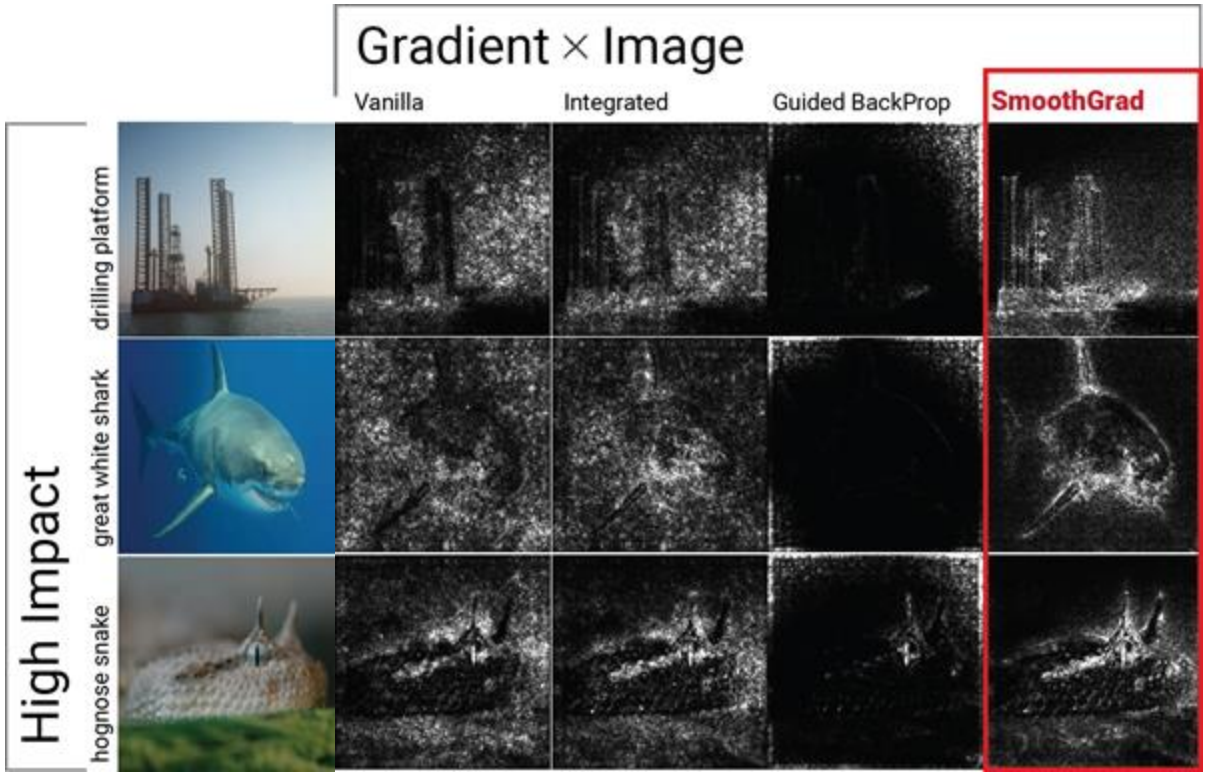
Experiments - Parameters



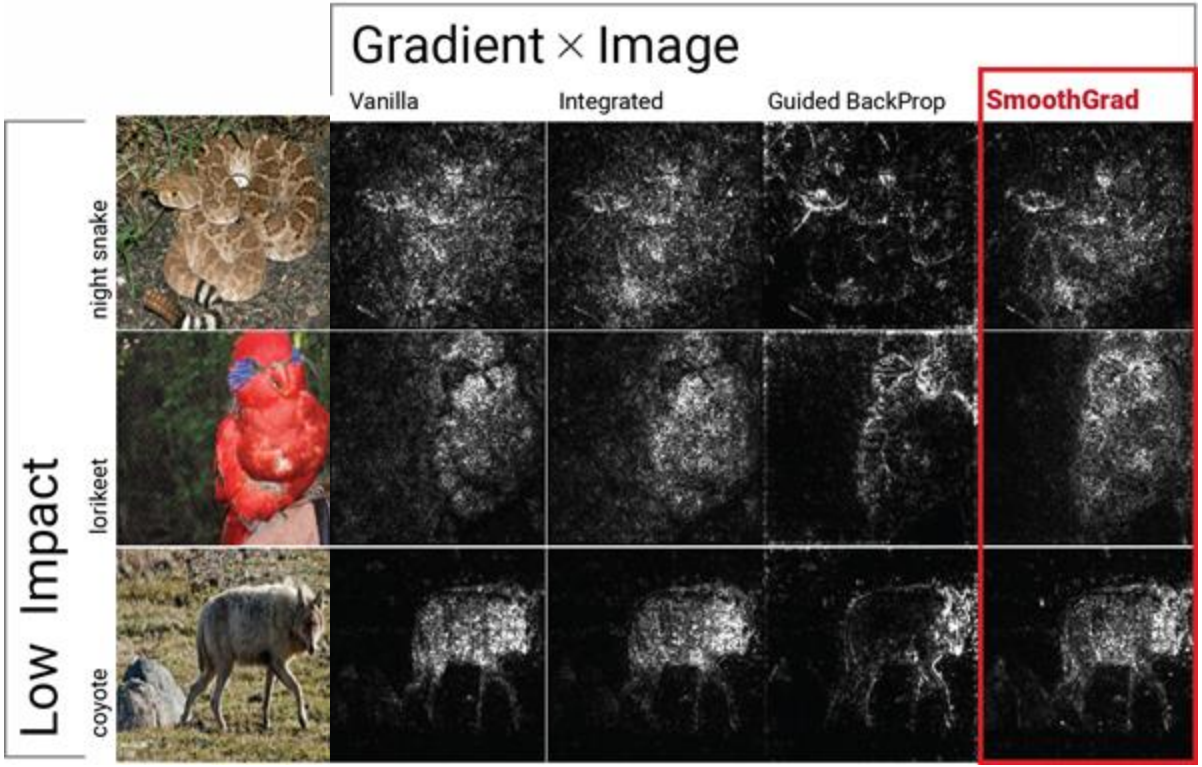
Experiments - Comparison to Baseline Methods

- Previous work is compared, as there is no ground truth for evaluation
- The Vanilla sensitivity map and other methods are used for comparison
- SmoothGrad was found to have more impact when an image had an object surrounded by a uniform background
- SmoothGrad was found to be more coherent than the Integrated and Vanilla sensitivity maps

Experiments - Comparison to Baseline Methods

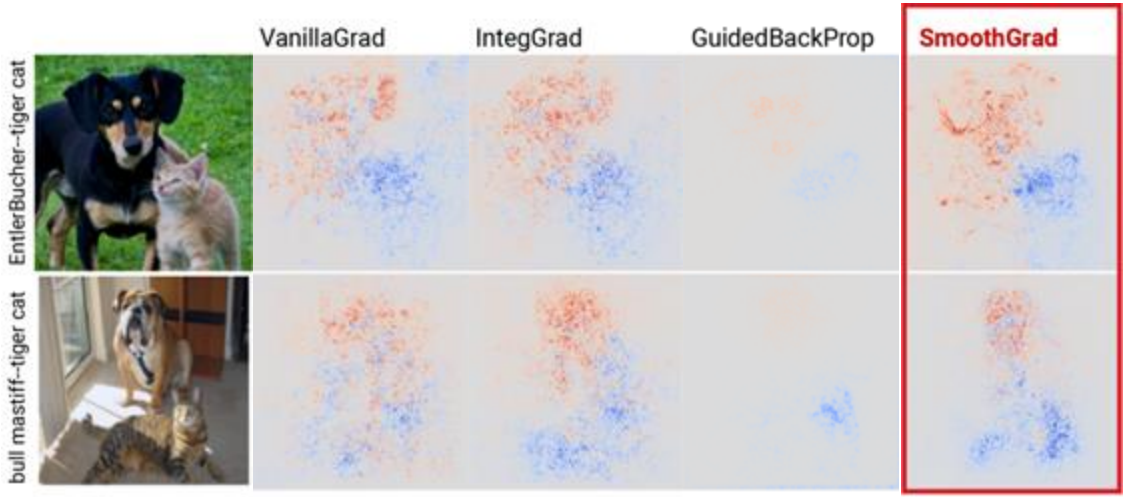


Experiments - Comparison to Baseline Methods



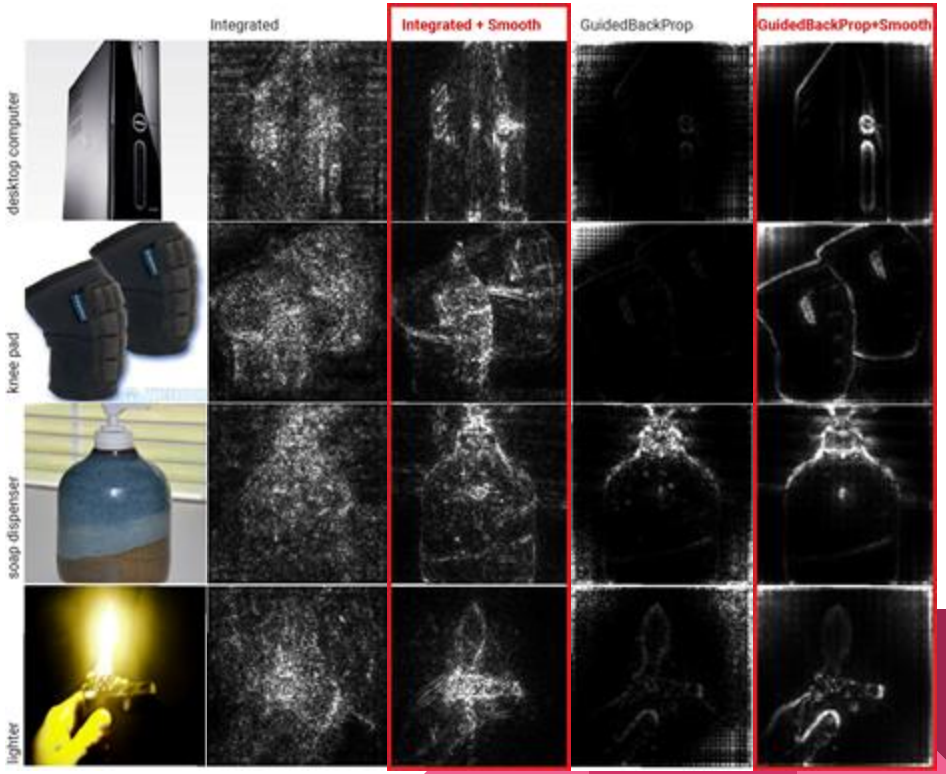
Experiments - Comparison to Baseline Methods

- Discriminativity of SmoothGrad is compared to the other methods
- Images with two objects of different classes were used
- Difference of the sensitivity maps were used to create a diverging color map



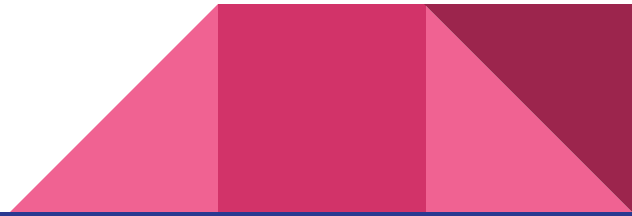
Experiments - Combining SmoothGrad

- SmoothGrad can be used to augment any other method
- SmoothGrad improves both the Integrated and Guided Backpropagation methods

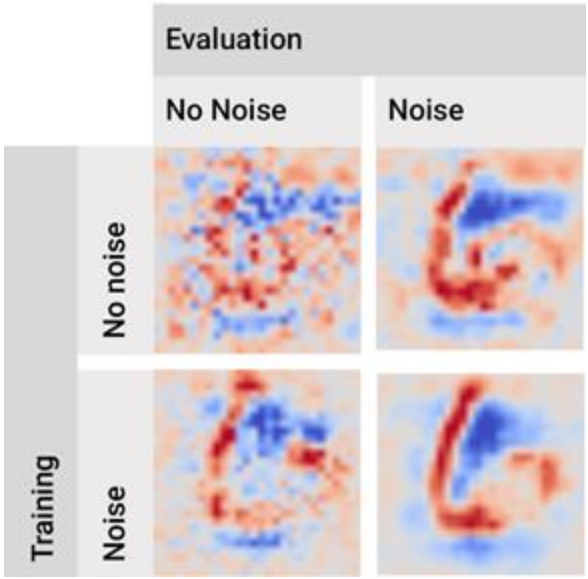


Experiments - Adding Noise During Training

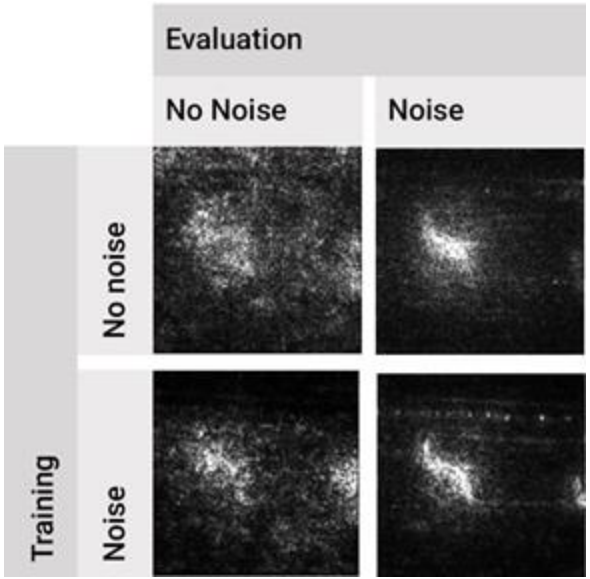
- SmoothGrad can be considered a regularization technique
- Applying SmoothGrad to samples during training was found to improve the sharpness of the sensitivity map
- Furthermore, training and inferring with SmoothGrad can have an additive effect making the sensitivity maps more coherent



Experiments - Adding Noise During Training



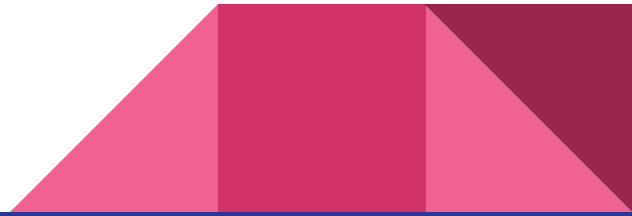
Effect of Adding Noise During Training vs Evaluation for MNIST



Effect of Adding Noise During Training vs Evaluation for Inception v3

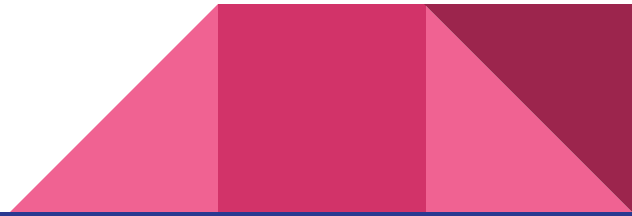
Conclusion

- Sensitivity maps visualize important features of a classifier’s predictions
- Sensitivity maps are noisy due to varying local gradients
- SmoothGrad visually sharpens sensitivity maps better than previous gradient-based techniques by:
  - Averaging maps made from many small perturbations of a given image
  - Training on data that has been perturbed with random noise



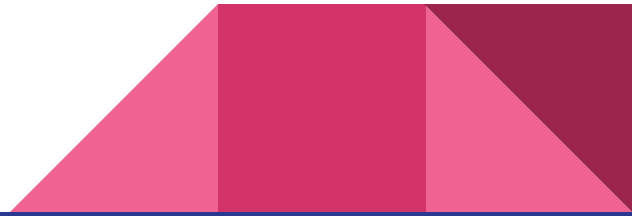
For Paper

- SmoothGrad visually sharpens sensitivity maps better than previous gradient-based techniques
- SmoothGrad is simple in concept and relatively easy to implement
- Researchers propose plausible explanation for noise in sensitivity maps



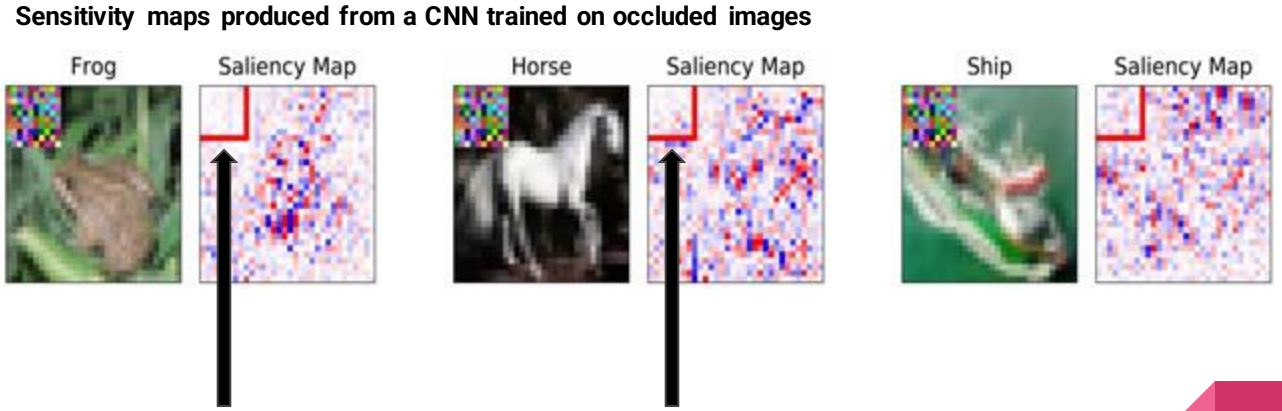
Against Paper

- Many ideas are not novel
- Raw gradients for sensitivity maps not new (e.g. Simonyan et al., 2013)
- Adding noise during training is a common regularization technique (Bishop, 1995)
- Stochastic approximations have solved intractable problems in statistics, economics, physics and machine learning



Against Paper

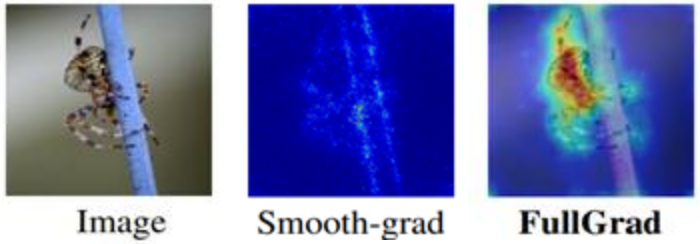
- Authors' hypothesis that sensitivity maps are truthful depictions of what network is doing has been disputed (Kim, Seo, et al., 2019)



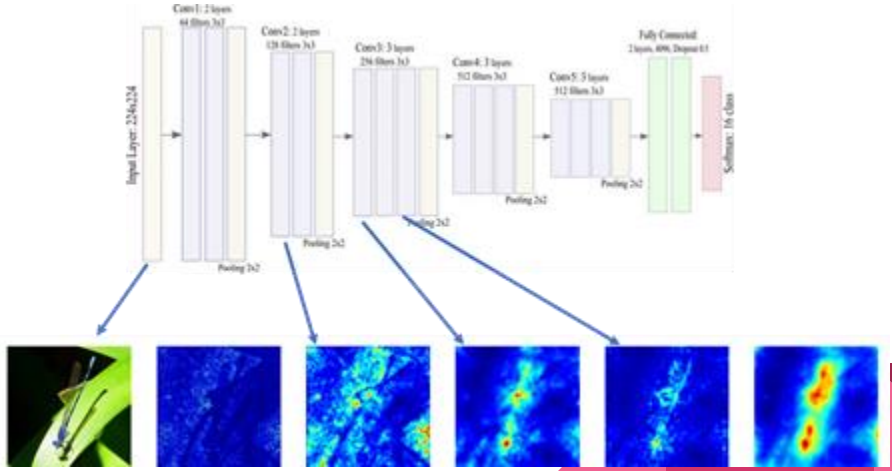
Sensitivity maps are non-zero on occluded patch despite patch being irrelevant to object

Against Paper

- Still produces noisy object boundaries
  - Later works segment and highlight salient regions (Srinivas et al., 2019)



VGG16



Any Questions?