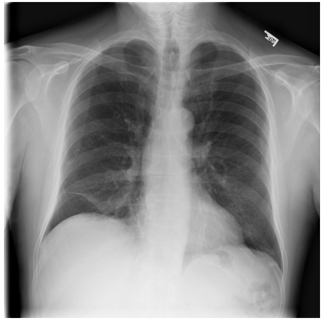
Full-Gradient Representation for Neural Network Visualization

Suraj Srinivas Francois Fleuret
Idiap Research Institute & EPFL

Why Interpretability for Deep Learning?



Deep
Neural
Network

Pneumonia

Why does the model think this chest x-ray shows signs of pneumonia?

Required for human-in-the-loop decision-making

Why Interpretability for Deep Learning?

Deep
Neural
Network

Gray Whale

Why does the model think this is a gray whale?

Required for human engineers to build better models

Saliency Maps for Interpretability

Deep
Neural
Network

Saliency
Algorithm

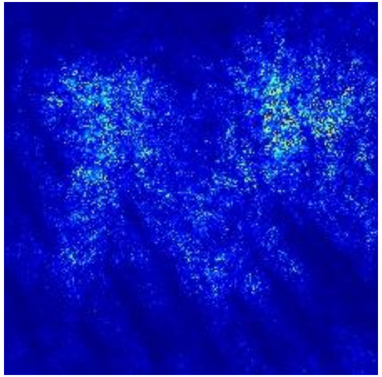
Highlight important regions

But what is “importance”?

Input-gradients for Saliency

Input - x

Saliency map - S



Neural network

$$y = f(x)$$

$$S = \nabla_x f(x)$$

- Clear connection to neural network function
- Saliency maps can be noisy and 'uninterpretable'

Simonyan et. al, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013

Wild West of Saliency Algorithms

1. Input-Gradients
2. Guided Backprop
3. Deconvolution
4. Grad-CAM
5. Integrated gradients
6. DeepLIFT
7. Local Relevance Propagation
8. Deep Taylor Decomposition

There is no single formal definition of saliency / feature importance accepted in the community.

Two Broad notions of Importance

- **Local importance** (Weak dependence on inputs)

  “A pixel is important if slightly changing that pixel,
  drastically affects model output”

- **Global importance** (Completeness with a baseline)

  “All pixels contribute numerically to the model output. The importance of a pixel is the extent of its contribution to the output.”
  E.g.: output = (contributions of) pixel1 + pixel2 + pixel3

The Nature of Importances

Still able to recognise bird

??

Sum of importances of pixels in the group ≠ Importance of group of pixels

https://pixabay.com/photos/kingfisher-bird-blue-plumage-1905255/

An Impossibility Theorem

For any piecewise linear function, it is **impossible** to obtain a saliency map that satisfies both **weak dependence** and **completeness with a baseline**.

Why? Saliency maps are **not expressive enough** to capture the complex non-linear interactions within neural networks.

Full-Gradients

Full-Gradients

For any neural network *f*(.) the following holds locally:

$$f(\mathrm{x}; w, b) = \nabla_x f(\mathrm{x}; w, b)^T \mathrm{x} + \nabla_b f(\mathrm{x}; w, b)^T b$$

Input sensitivity

Neuron sensitivity
(Gradients w.r.t.
intermediate activations)

x: input

w: weights

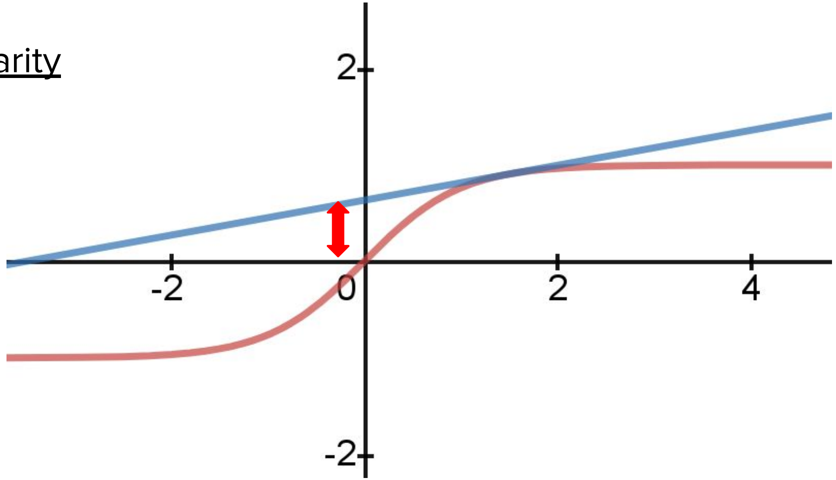
b: biases
concatenated
across layers

Neural Network Biases

Batch Normalization

$$(\frac{\mathrm{x}-\mu}{\sigma})\times w+b=\frac{\mathrm{x}\times w}{\sigma}-\frac{\mu\times w}{\sigma}+b$$

Non-linearity



Local linear approximation

y = tanh(x)

Properties of Full-gradients

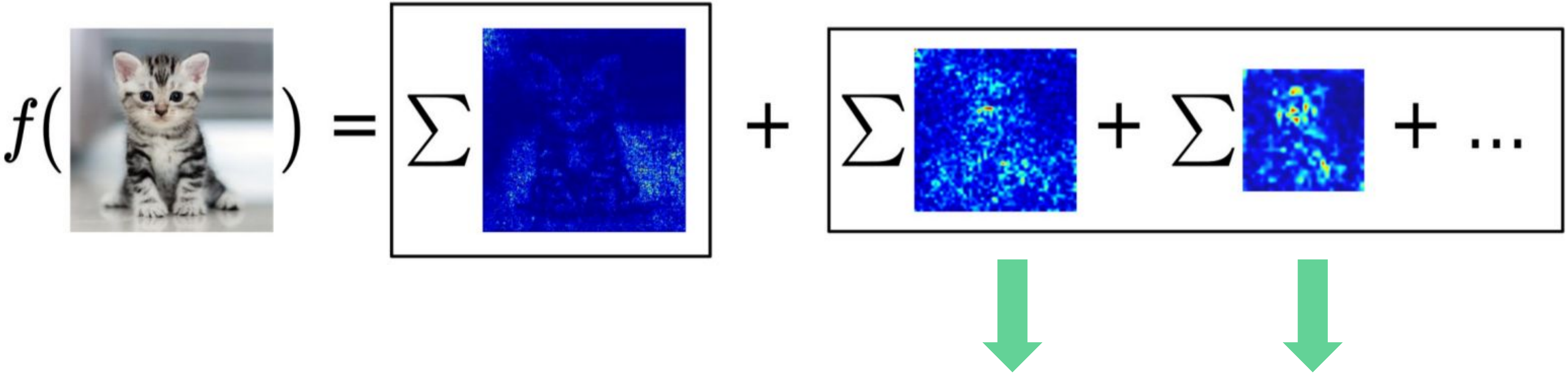
- Satisfies both **weak dependence** and **completeness with a baseline**, since full-gradients are more expressive than saliency maps

- Does not suffer from non-attribution due to **saturation.** Many input-gradient methods provide zero attribution in regions of zero gradient.

- **Fully sensitive** to changes in underlying function mapping. Some methods (e.g.: guided backprop) do not change their attribution even when some layers are randomized.

Adebayo et. al,. Sanity Checks for Saliency Maps, 2018

Full-Gradients for Convolutional Nets



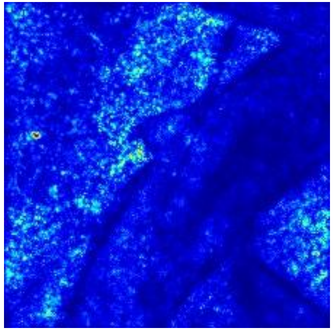
bias-gradients of neurons in layer 1

bias-gradients of neurons in layer 2

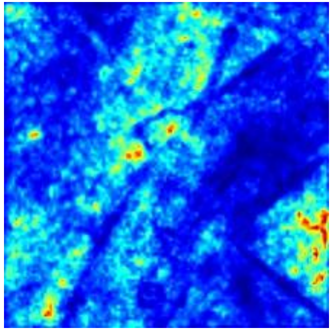
Naturally incorporates **importance** of a pixel at **multiple** receptive fields!

FullGrad Aggregation

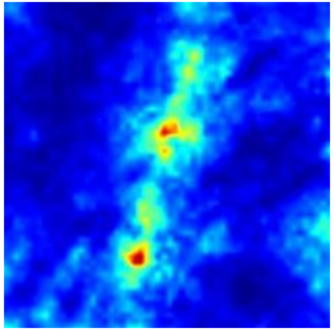
Image



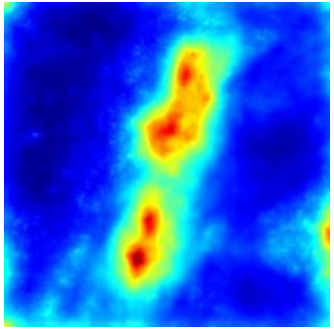
Input-gradients



Bias-gradients layer 3



Bias-gradients layer 5



FullGrad Aggregate

$$S(\mathrm{x}) = \psi(\nabla_x f(\mathrm{x}; b) \odot \mathrm{x}) + \sum_{l}^{\#\text{layers}} \sum_{c}^{\#\text{channels}} \psi(\nabla_b f(\mathrm{x}; b)_{l,c} \odot b_{l,c})$$

FullGrad Saliency Maps

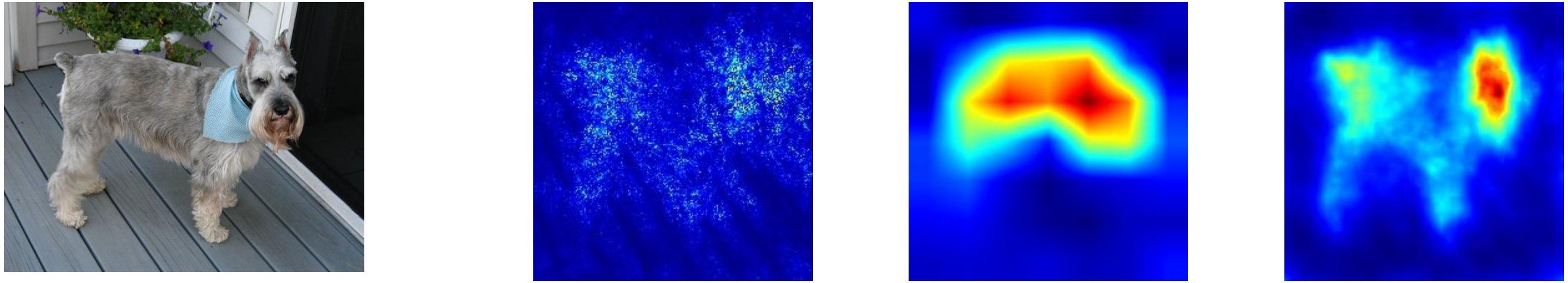
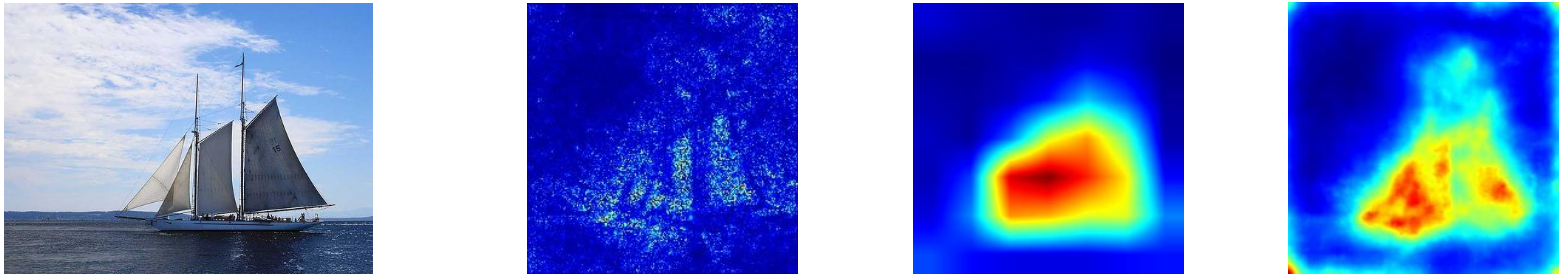
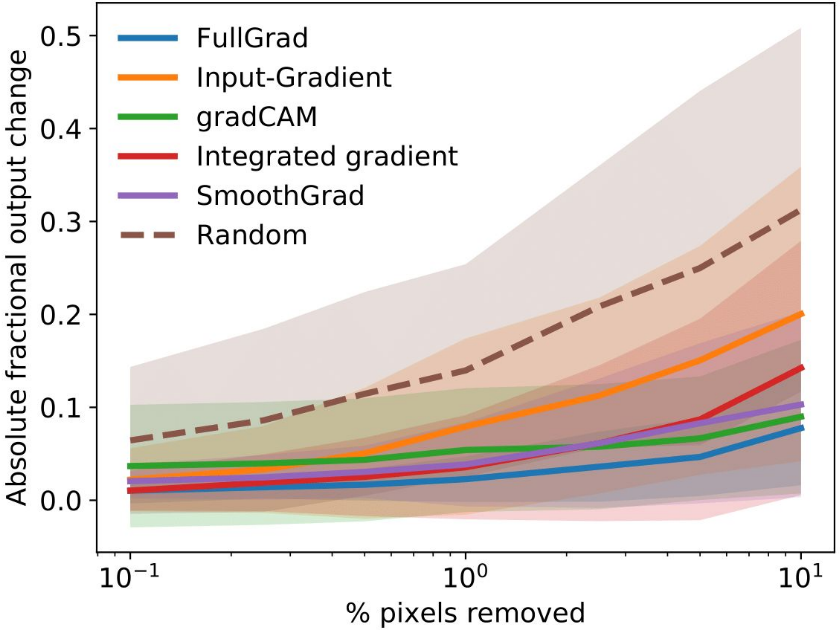


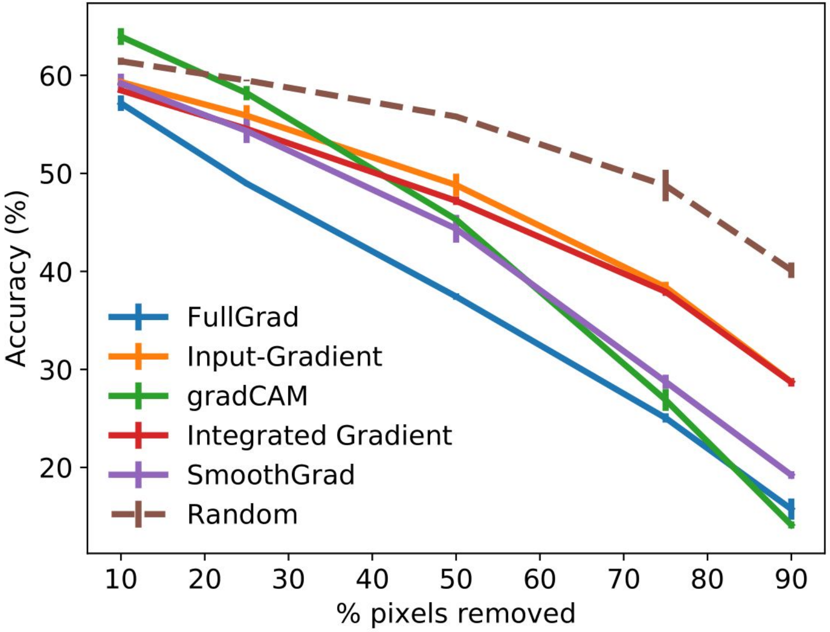
Image | Input-gradients | Grad-CAM | **FullGrad (Ours)**



Quantitative Results



Pixel perturbation test



Remove and Retrain (ROAR) test

Conclusion

- We have introduced a new tool called **full-gradient representation** useful for visualizing neural network responses

- For convolutional nets, **FullGrad** saliency map naturally captures the importance of a pixel at multiple scales / contexts

- **FullGrad** better identifies important image pixels than other methods

Code: https://github.com/idiap/fullgrad-saliency

Thank you