Full-Gradient Representation for Neural Network Visualization

Suraj Srinivas & François Fleuret
Idiap Research Institute & EPFL

Saliency maps capture **importance** of an input part for a specific **task** performed by a neural network. How is such **importance** defined?

Local vs Global importance

There are two different notions of saliency used in literature

- **Local** importance captures model sensitivity to input
- **Global** importance captures the ability to recover model output using the saliency map (a.k.a. *completeness* of saliency map)

Question: Can a saliency method satisfy both these properties?
Answer: No. (Proposition 1 in the paper)
Why? Saliency methods are too restrictive.
Implications: One can always find **counter-intuitive** behaviour for saliency maps by violating some notion of importance.

Full-Gradients

$f(\cdot) \rightarrow$ neural network, $\mathbf{x} \rightarrow$ input
$\mathbf{w} \rightarrow$ weights of all layers, $\mathbf{b} \rightarrow$ biases of all layers

$$f(\mathbf{x};\mathbf{w},\mathbf{b}) = \underbrace{\nabla_{\mathbf{x}} f(\mathbf{x};\mathbf{w},\mathbf{b})^T \mathbf{x}}_{\text{input-gradients}} + \underbrace{\nabla_{\mathbf{b}} f(\mathbf{x};\mathbf{w},\mathbf{b})^T \mathbf{b}}_{\text{bias-gradients}}$$

Bias-gradients $\rightarrow$ gradient of output w.r.t. intermediate features

Full-gradients satisfy both notions of importance as they are more expressive than saliency maps.

FullGrad Saliency

We propose **FullGrad** saliency which incorporates both input-gradients and feature-level bias-gradients.

$$S_f(\mathbf{x}) = \psi(\nabla_{\mathbf{x}} f(\mathbf{x};\mathbf{w},\mathbf{b}) \odot \mathbf{x}) + \sum_{\texttt{layers}} \sum_{\texttt{channels}} \psi\left(\nabla_{\mathbf{b}} f(\mathbf{x};\mathbf{w},\mathbf{b}) \odot \mathbf{b}\right)$$

where $\psi(\cdot)$ is a normalization function.

We show that any neural network's output score can be decomposed into an **input-gradient** term and **per-neuron** gradient terms.

For ConvNets, we find that **aggregating** these gradient maps lead to improved saliency maps.

arXiv

Visualizations

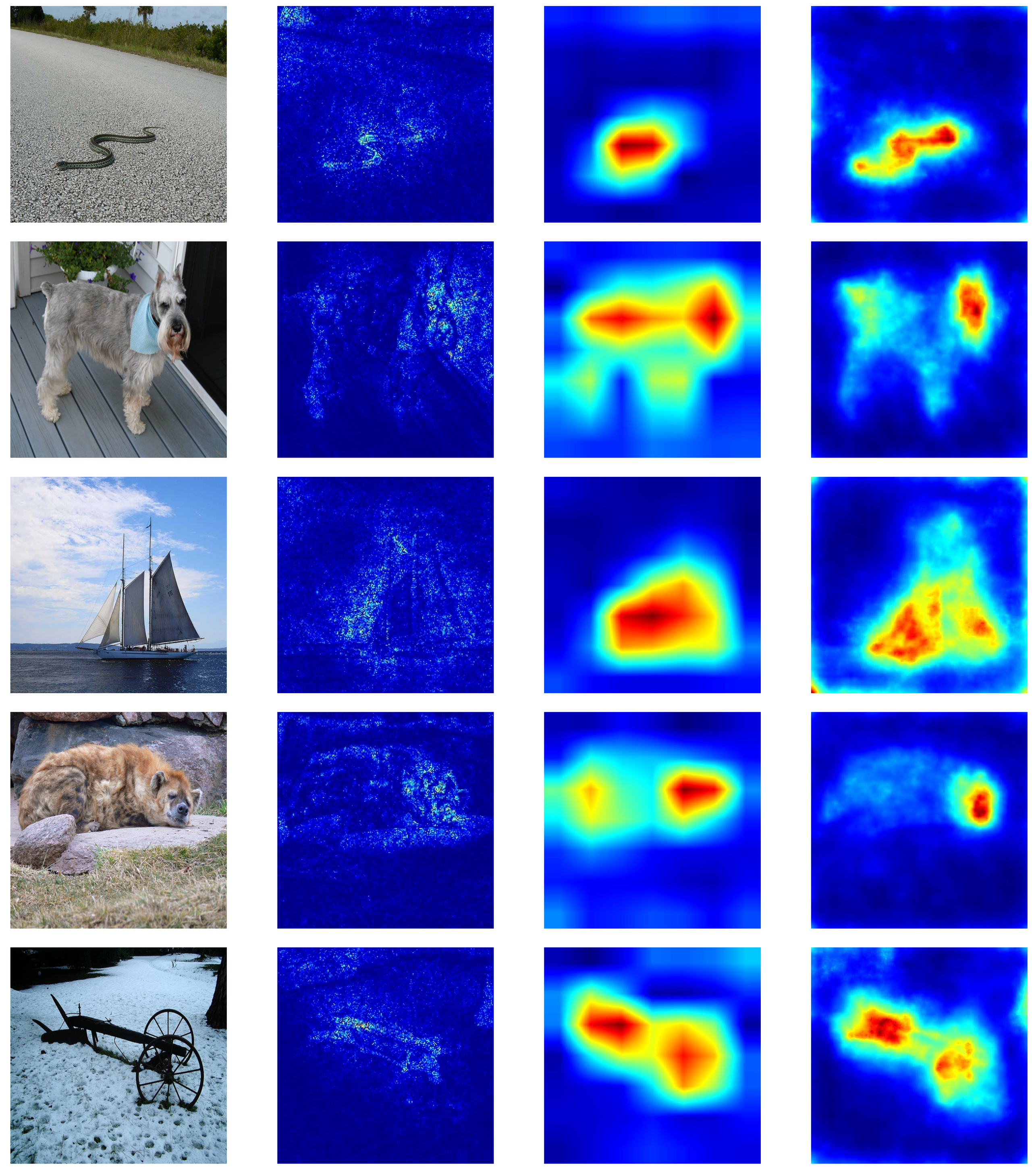
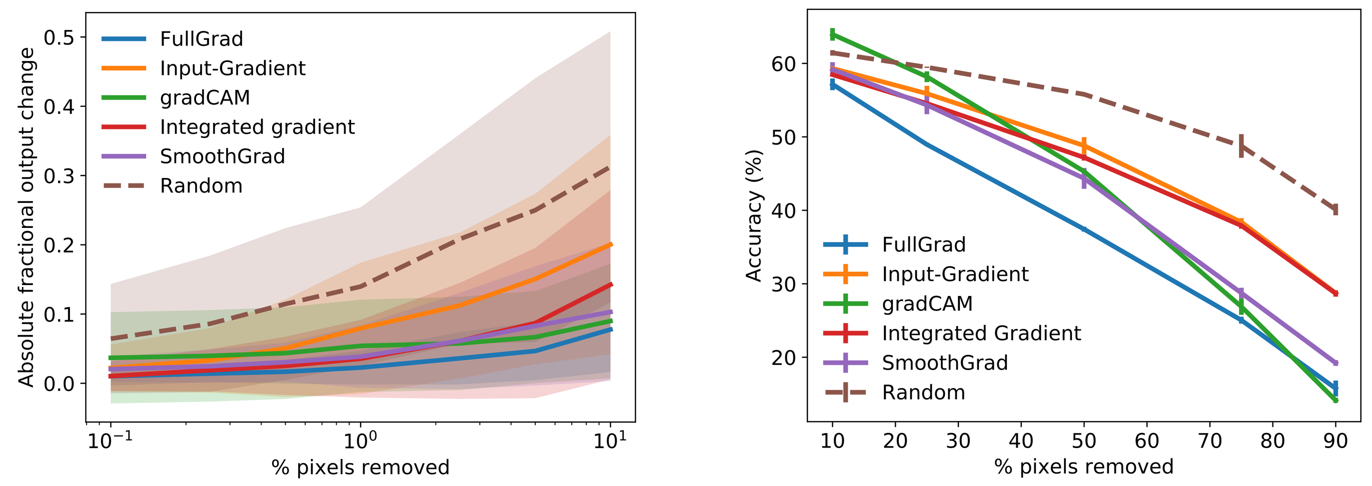


Image | Integrated gradient [1] | Grad CAM [2] | **FullGrad (Ours)**

Experiments



(Left) Pixel sensitivity test: remove least salient pixels and observe change in output. Smaller is better. **(Right)** Remove and Retrain (ROAR) test: remove most salient pixels in training set, retrain, and observe accuracy. Smaller is better.

References

[1] Sundararajan et. al., "Axiomatic attribution for deep networks", 2017
[2] Selvaraju et. al., "Grad-cam: Visual explanations from deep networks via gradient-based localization", 2017