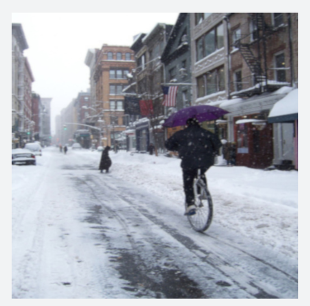
Rethinking the Role of Gradient-Based Attribution Methods for Model Interpretability

Suraj Srinivas & François Fleuret
suraj.srinivas@idiap.ch, francois.fleuret@unige.ch

Saliency Maps represent "importance" of input pixels as seen by a model performing a task.


Input Image


Saliency Map

Research Question: Why are saliency maps highly structured and interpretable for standard models, even when this is not enforced during training?

Our Findings:

- Gradient-based saliency maps can be arbitrary due to the shift-invariance of softmax, even for models that generalize perfectly
- Structure of gradient-based saliency maps depends on the class-conditional generative model $\mathbf{p}(\mathbf{x} \mid \mathbf{y})$ and not the discriminative model $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$, which they are used to interpret
- Improving this generative model using score-matching improves gradient interpretability, while deteriorating this generative model has the opposite effect

Implicit Density Models

The logits $f_i(x)$ of softmax-based models for class i...

$$p_\theta(y = i \mid x) = \frac{\exp f_i(x)}{\sum_j \exp f_j(x)} = \frac{p_\theta(x \mid y = i)p(y = i)}{\sum_j p_\theta(x \mid y = j)p(y = j)}$$

...can be viewed as an energy function ...

$$p_\theta(x \mid y = i) = \frac{\exp f_i(x)}{Z}$$

...and the logit-gradients as gradients of the log density!

$$\boxed{\underbrace{\nabla_x \log p_\theta(x \mid y = i)}_{\text{gradients of log density}} = \underbrace{\nabla_x f_i(x)}_{\text{logit-gradients}}}$$

This leads to the following hypothesis:

> Logit-gradients are highly structured because of their alignment with the ground truth gradients
>
> $\nabla_x \log p_{data}(x \mid y) \approx \nabla_x \log p_\theta(x \mid y) = \nabla_x f_i(x)$

Score-Matching

Score-Matching is a generative modelling principle based on aligning gradients of log density, by minimizing the following objective.

$$J(\theta) = \mathbb{E}_{p_{data}(\mathbf{x})} \frac{1}{2} \|\nabla_x \log p_\theta(\mathbf{x}) - \nabla_x \log p_{data}(\mathbf{x})\|_2^2$$

This can be re-written as an objective which omits the unknown $\nabla_x \log p_{data}(x)$ term.

$$J(\theta) = \mathbb{E}_{p_{data}(\mathbf{x})} \left(\text{trace}(\nabla_x^2 \log p_\theta(\mathbf{x})) + \frac{1}{2} \|\nabla_x \log p_\theta(\mathbf{x})\|_2^2 \right) + C$$

Interpretability $\iff$ Generative Modelling

Implicit density modelling perspective reveals generative modelling interpretations of the following methods ordinarily used for interpretability.

- Logit-gradients $\iff$ gradients of log density
- Activation maximization of logits $\iff$ MCMC sampling via Langevin dynamics
- Pixel perturbation test $\iff$ density ratio test

Experimental Setup

Objective: Train models with different levels of gradient alignment by regularization, and study their effect on input-gradient interpretability.

$$\underbrace{\ell_{reg}(f(\mathbf{x}), i)}_{\text{regularized loss}} = \underbrace{\ell(f(\mathbf{x}), i)}_{\text{cross-entropy}} + \lambda \underbrace{R(\mathbf{x})}_{\text{Regularizer}}$$

Regularized Score-Matching: We propose relaxations of score-matching to overcome computational intractability of Hessian trace estimation, and stabilize the objective, which we use as a regularizer.

$$h(\mathbf{x}) := \frac{2}{\sigma^2} \mathbb{E}_{v \sim \mathcal{N}(0, \sigma^2 I)} \left(f_i(\mathbf{x} + v) - f_i(\mathbf{x}) \right)$$

$$R(\mathbf{x}) = \left(\underbrace{\overbrace{h(\mathbf{x})}^{\text{Hessian-trace}} + \frac{1}{2} \overbrace{\|\nabla_x f_i(\mathbf{x})\|_2^2}^{\text{gradient-norm}}}_{\text{score-matching}} + \underbrace{\overbrace{\mu}^{10^{-4}} h^2(\mathbf{x})}_{\text{stability regularizer}} \right)$$

Other Baselines: We use the following other baseline regularizations for comparison

- No regularization
- Anti-score-matching regularization, where hessian trace is maximized instead of being minimized
- Gradient norm regularization, where norm of input-gradients is minimized

Effect on Generative Modelling

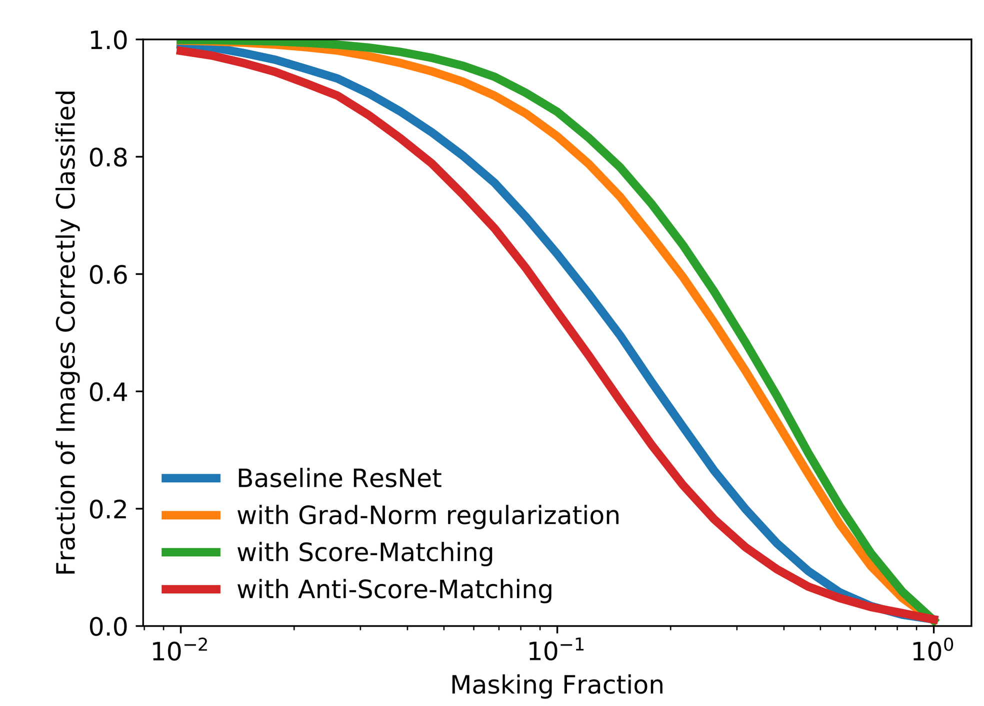
We measure sample quality using the GAN-test scores (higher is better) on samples generated from the implicit density models via Langevin MCMC.

Model	GAN-test (%)
Baseline ResNet	59.47
+ Anti-Score-Matching	16.40
+ Gradient Norm-regularization	**80.07**
+ Score-Matching	**72.75**

Conclusion: Score-matching and gradient-norm regularization improves, while anti-score-matching deteriorates sample quality.

Effect on Interpretability

We measure a proxy for interpretability using the pixel perturbation test (higher is better) on the input-gradients of various models.



Conclusion: Score-matching and gradient-norm regularized models improves, while anti-score-matching deteriorates gradient interpretability.

Acknowledgements

FNSNF