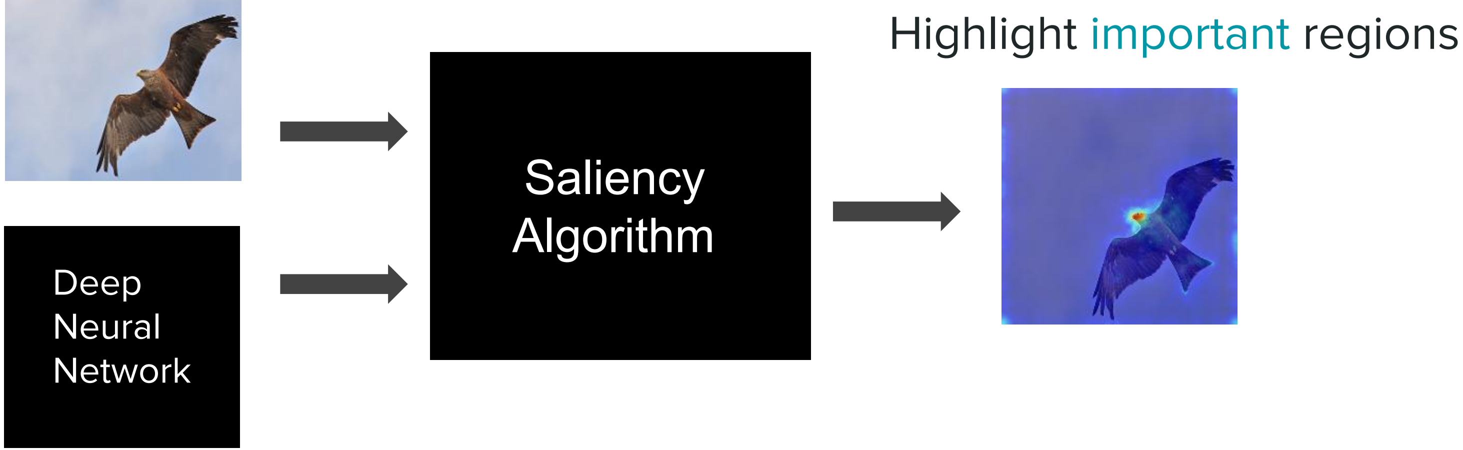
Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability

Suraj Srinivas[1] & François Fleuret[2]

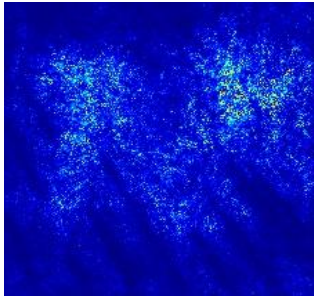
Idiap Research Institute[1] & EPFL[1], University of Geneva[2]

Saliency Maps for Model Interpretability



Input-gradient Saliency

Input (x)



Saliency map (S)

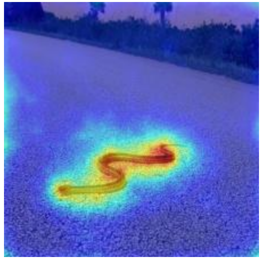
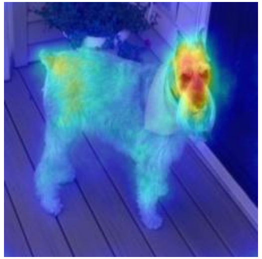
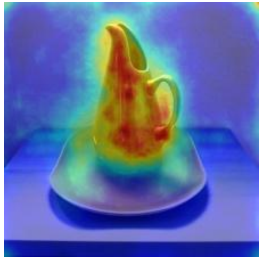
Neural network

$$y = f(x)$$

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

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

Why are gradients highly structured anyway?

Gradient Structure is Arbitrary

$$s_i(\mathrm{x}) = \frac{\exp(f_i(\mathrm{x}))}{\sum_{j=1}^{C} \exp(f_j(\mathrm{x}))} = \frac{\exp(f_i(\mathrm{x}) + g(\mathrm{x}))}{\sum_{j=1}^{C} \exp(f_j(\mathrm{x}) + g(\mathrm{x}))}$$

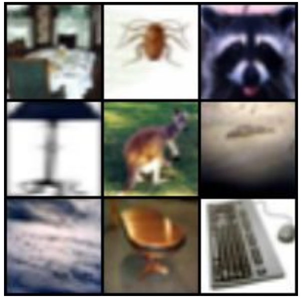
Arbitrary!

$$\tilde{f}_i(\mathrm{x}) = f_i(\mathrm{x}) + g(\mathrm{x}) \implies \nabla_x \tilde{f}_i(\mathrm{x}) = \nabla_x f_i(\mathrm{x}) + \nabla_x g(\mathrm{x})$$

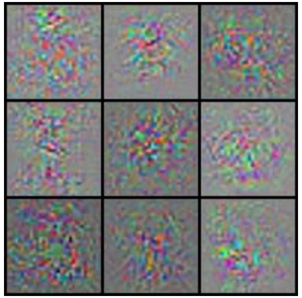
Pre-softmax (logit) gradients can be arbitrary, even if the model generalizes perfectly!

This also holds for post-softmax gradients (see paper for details).

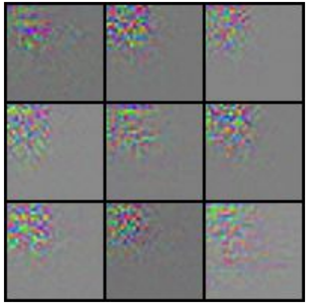
Gradient Structure is Arbitrary



Input image



Logit-gradients of standard model

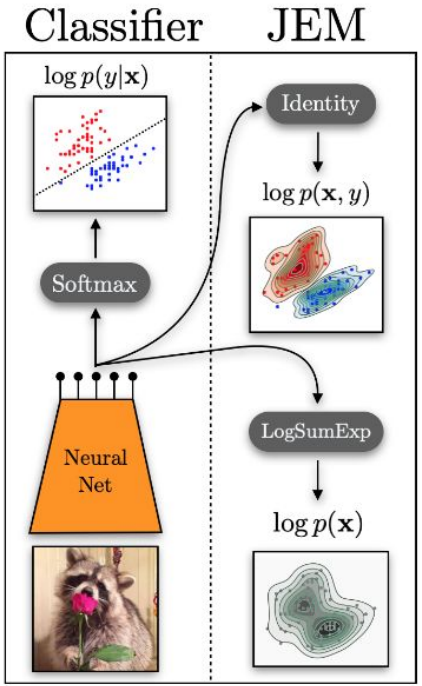


Logit-gradients of model with “fooled” gradients

Logit gradients don’t need to encode relevant information, but they still do. Why?

Generative Models hidden within Discriminative Models

Implicit Density Models within Discriminative Models



$$p(y = i \mid \mathrm{x}) = \frac{\exp(f_i(\mathrm{x}))}{\sum_{j=1}^{C} \exp(f_j(\mathrm{x}))} = \frac{p(\mathrm{x} \mid y = i)p(y = i)}{\sum_{j=1}^{C} p(\mathrm{x} \mid y = j)p(y = j)}$$

$$p(\mathrm{x} \mid y = i) = \frac{\exp f_i(\mathrm{x})}{\int_{x'} \exp f_i(\mathrm{x}')}$$

$$\nabla_x \log p(\mathrm{x} \mid y = i) = \nabla_x f_i(\mathrm{x})$$

Grathwohl et. al, Your Classifier is Secretly an Energy-based Model and You Should Treat it Like One, ICLR 2020

Hypothesis

$$\nabla_x f_i(\mathrm{x}) = \nabla_x \log p_\theta(\mathrm{x} \mid y = i) \approx \nabla_x \log p_{data}(\mathrm{x} \mid y = i)$$

Hypothesis: The structure of logit-gradients is due to its alignment with the ground truth gradients of log density.

A concrete test: Increasing gradient alignment must improve gradient interpretability & decreasing this alignment must deteriorate interpretability.

Training Energy-based Models

Energy-based Generative Models

$$p(\mathrm{x} \mid y = i) = \frac{\exp f_i(\mathrm{x})}{\int_{x'} \exp f_i(\mathrm{x}')}$$

- Sampling via MCMC: Use Langevin Dynamics ("noisy gradient ascent")

$$\mathbf{x}_0 \sim p_0(\mathbf{x}), \qquad \mathbf{x}_{i+1} = \mathbf{x}_i - \frac{\alpha}{2}\frac{\partial E_\theta(\mathbf{x}_i)}{\partial \mathbf{x}_i} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \alpha)$$

- Training using:
  - *Approx. Max-likelihood* - requires MCMC to estimate normalizing constant
  - *Score-matching* - does not require normalizing constant, but is unstable
  - *Noise Contrastive Estimation*, Minimizing *Stein Discrepancy*, etc

Score-Matching

Alignment of gradients is a generative modelling principle!

$$
\begin{aligned}
J(\theta) &= \mathbb{E}_{p_{data}(\mathbf{x})} \frac{1}{2} \|\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{data}(\mathbf{x})\|_2^2 \\
&= \mathbb{E}_{p_{data}(\mathbf{x})} \left(\text{trace}(\nabla^2_{\mathbf{x}} \log p_\theta(\mathbf{x})) + \frac{1}{2} \|\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})\|_2^2 \right) + \texttt{const}
\end{aligned}
$$

→ Does not require $\nabla_x \log p_{data}(x)$

- Hessian computation is intractable for deep models!
- Trace of Hessian is unbounded below

Aapo Hyvarinen. “Estimation of non-normalized statistical models by score matching”.Journal of Machine Learning Research, 6(Apr):695–709, 2005

Regularized Score-Matching

Efficient estimation of Hessian trace

$$tr(\nabla_x^2 \log p(x)) = \mathbb{E}_{v\sim\mathcal{N}(0,I)}\ v^\top \nabla_x^2 \log p(x) v$$ ⟶ Hutchinson's trick

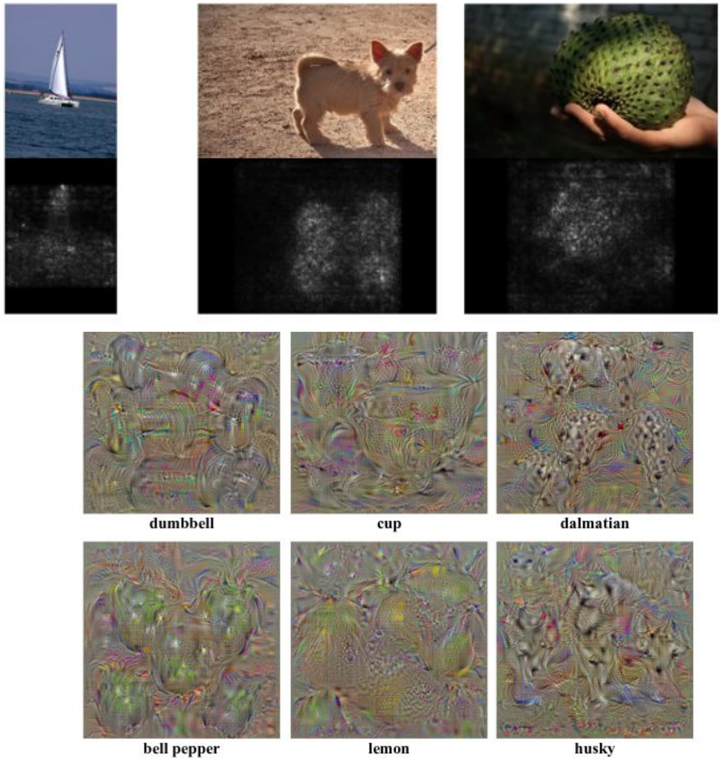
$$\approx \frac{2}{\sigma^2}\mathbb{E}_{v\sim\mathcal{N}(0,\sigma^2 I)}(\log p(x+v) - \log p(x))$$ ⟶ Taylor series

Regularization of Hessian trace

$$J(\theta) = tr(\nabla_x^2 \log p_\theta(x)) + \frac{1}{2}\|\nabla_x \log p(x)\|^2 + \overbrace{\mu}^{10^{-4}}\ (tr(\nabla_x^2 \log p_\theta(x)))^2$$

Interpretability vs Generative Modelling

Interpretability	Generative Modelling
Logit-Gradients	Gradient of log p(x)
"Deep dream" Visualization by Activation Maximization	MCMC Sampling by Langevin Dynamics
Pixel perturbation test	Density ratio test



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

Experiments

We compare generative capabilities and gradient interpretability across different models

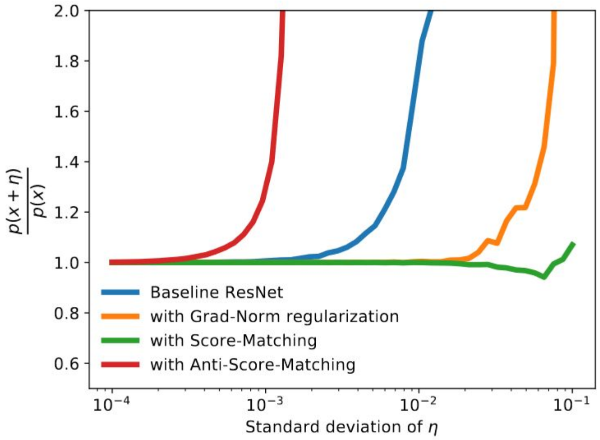
- Baseline unregularized model
- Score-matching regularized model
- Anti-score-matching regularized model
- Gradient norm regularized model

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

$$\underbrace{\ell_{reg}(f(\mathbf{x}), i)}_{\text{regularized loss}} = \underbrace{\ell(f(\mathbf{x}), i)}_{\text{cross-entropy}} + \lambda \left(\underbrace{\overbrace{h(\mathbf{x})}^{\text{Hessian-trace}} + \frac{1}{2}\overbrace{\|\nabla_{\mathbf{x}} f_i(\mathbf{x})\|_2^2}^{\text{gradient-norm}}}_{\text{score-matching}} + \underbrace{\overbrace{\mu}^{10^{-4}} h^2(\mathbf{x})}_{\text{stability regularizer}} \right)$$

Effect on Generative Modelling

$$\frac{p(x+\eta)}{p(x)} = \exp(f(x+\eta) - f(x))$$



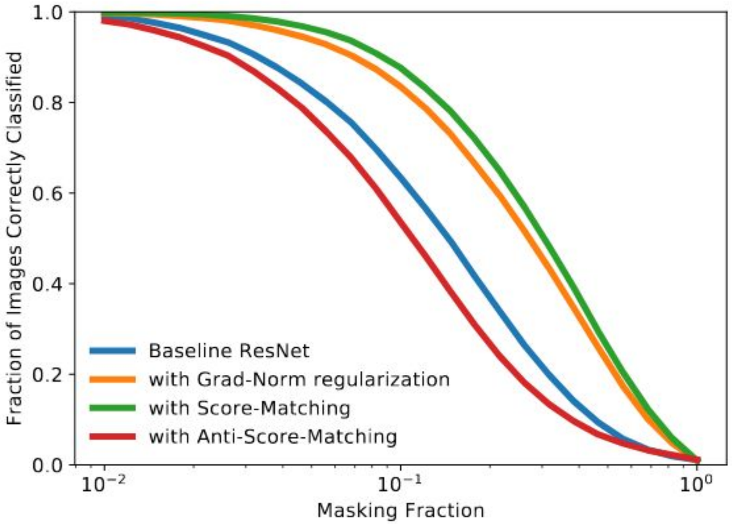
- Models assign high likelihoods to noisy points!
- This tendency reduces with score-matching models, and increases for anti-score-matching models

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

- Sample quality is measured using GAN-test
- Sample quality improves with score-matching and deteriorates with anti-score-matching

Schmelkov et. al, How good is my GAN?, ECCV 2018

Effect on Gradient Interpretability

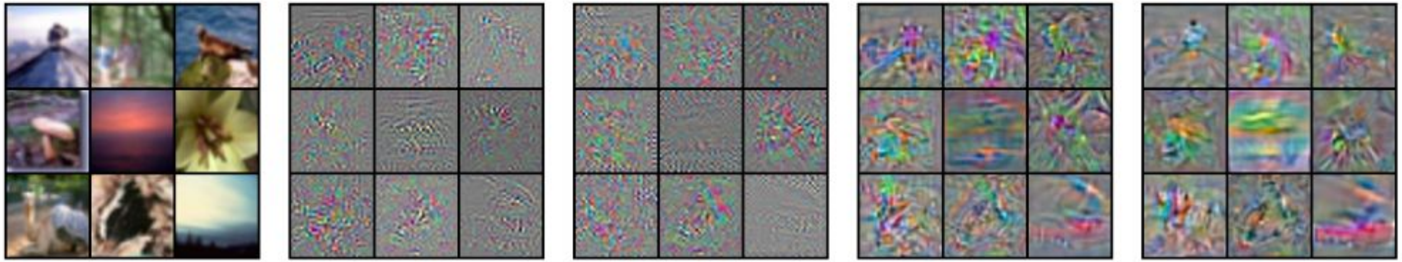


- A proxy for gradient interpretability is the pixel perturbation test, which masks unimportant pixels and checks accuracy (higher is better)

- Score-matching improves on this metric, while anti-score-matching deteriorates

This confirms our hypothesis that the implicit density modelling influences gradient interpretability.

Effect on Gradient Interpretability



(a) Input Image (b) Baseline ResNet (c) With Anti score-matching (d) With Gradient-norm regularization (e) With Score-matching

Conclusion

- We present evidence that logit-gradient interpretability is strongly related to the underlying class conditional density model $p(x|y)$ and **not** $p(y|x)$, which they are typically used to interpret.

- Broad message: Gradient structure depends on factors outside the discriminative properties of the model.

- Open Question: What causes approximate energy-based training in standard models?