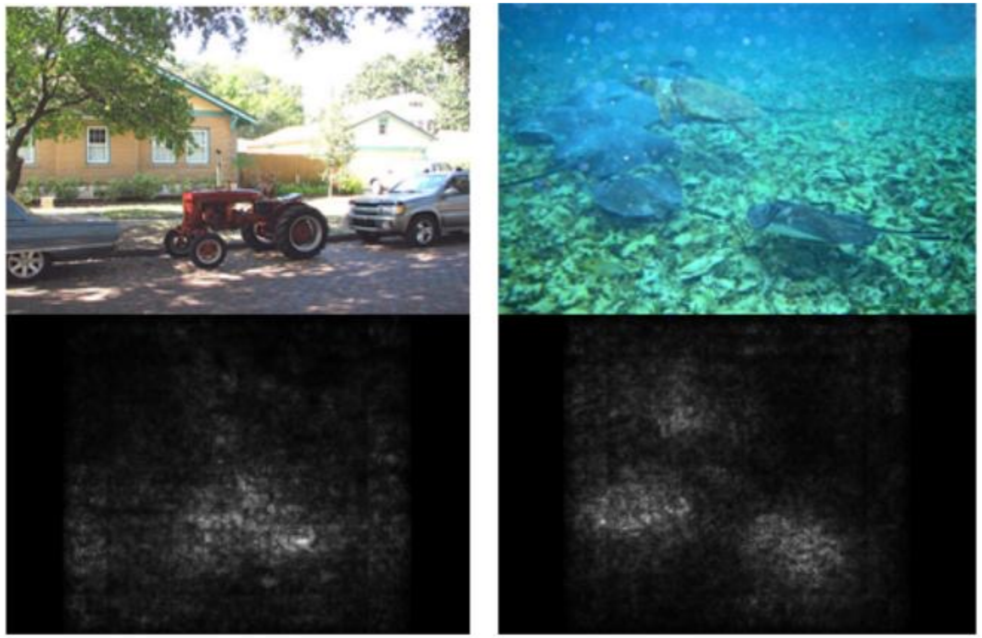
A Functional Information Perspective on Model Interpretation

Itai Gat, Nitay Calderon, Roi Reichart, Tamir Hazan

ICML 2022

Introduction

The most fundamental gradient-based explanation is the saliency map.



Introduction

- Prior works show that gradient-based interpretability methods could be noisy in part due to local variations in partial derivatives.
- In order to overcome this, prior work propose to compute the expected output of gradient-based methods with respect to their input.

Introduction

- SmoothGrad computes the expectation of the gradient under Gaussian perturbations:

$$\mathbb{E}_{z\sim\mathcal{N}(0,I)}[\nabla_x f(x+z)]$$

- Later, methods like SmoothGrad squared and VarGrad were proposed.
- Those methods are not backed up in theory.

Introduction

Issue: Existing sampling-based methods rely on the assumption that the features are uncorrelated.

In this work, we provide **a theoretical framework** that applies functional entropy as a guiding concept to the amount of information a given deep net holds for a given input with respect to any possible labels.

Background - notation

Let $f_y(x) = p_w(y|x)$ be the probability assigned to a label y by a model f on a data sample x.

The *functional entropy* of the non-negative label function $f_y \geq 0$ is

$$Ent_\nu(f_y) \triangleq \int_{\mathbb{R}^d} f_y(z) \log \frac{f_y(z)}{\int_{\mathbb{R}^d} f_y(z) d\mu(z)} d\nu(z).$$

Where $\nu = \mathcal{N}(x, \mathrm{I})$.

We hence define the functional entropy of a deep net with respect to a label y by the function Softmax output $f_y(z)$ when $z \sim \nu$.

Background - entropy and explainability

The functional entropy can be thought of as the KL divergence between the prior distribution $p_\nu(z)$ and the posterior distribution $q_\nu(z)$ of the decision function $f_y(z)$ with respect to the data generating distribution over z.

Then, we have:

$$Ent_\nu = KL\left(q_\nu(z)||p_\nu(z)\right).$$

Background - **Log-Sobolev inequality**

Instead of directly estimating the functional entropy (which is intractable), we use the log-Sobolev inequality.

This permits to bound the functional entropy with the *functional Fisher information*:

$$Ent_{\nu}\left(f_y\right) \leq \frac{1}{2}\mathcal{J}_{\nu}\left(f_y\right) \triangleq \frac{1}{2}\int_{\mathbb{R}^d} \frac{\left\langle \nabla f_y(z), \nabla f_y(z)\right\rangle}{f_y(z)}\, d\nu(z)\,.$$

Feature Contribution via Functional Fisher Information

We propose a sampling-based method that can quantify the contribution of an input feature x_i to the decision function f_y:

$$\mathcal{J}_\nu(f_y) = \sum_i \mathbb{E}\left[\frac{\left(\nabla f_y(z)_i\right)^2}{f_y(z)}\right]$$

We need to overcome two challenges to use functional entropy and functional Fisher information as guiding concepts.

Feature Contribution via Functional Fisher Information

Challenge: Real-world data features are correlated.

Theorem 1: *For every non-negative function* f_y *and a Gaussian measure* $\mu \sim \mathcal{N}(x, \Sigma)$

$$Ent_\mu(f_y) \leq \frac{1}{2}\int_{\mathbb{R}^d} \frac{\langle \Sigma \nabla f_y(z), \nabla f_y(z) \rangle}{f_y(z)} d\mu(z)\,.$$

Feature Contribution via Functional Fisher Information

Challenge: Computation of subset of features

Theorem 2: *For a partitioned input $x = (x_1, x_2)$, a Gaussian measure μ, a conditional distribution μ_1, and a marginal distribution μ_2. For every non-negative function f_y: $\mathbb{R}^d \to \mathbb{R}$,*

$$Ent_\mu(f_y) \leq \frac{1}{2}\mathbb{E}_{z_2 \sim \mu_2}[\mathcal{J}_{\mu_1}(f_y|z_2)].$$

And,

$$Ent_{\mu_1}(f_y|x_2) \leq \frac{1}{2}\mathcal{J}_{\mu_1}(f_y|z_2).$$

Thanks for listening!

Code

Arxiv