On the (In)fidelity and Sensitivity of Explanations

Chao Chen @CSE Dept

Preliminaries – Notation in explanations

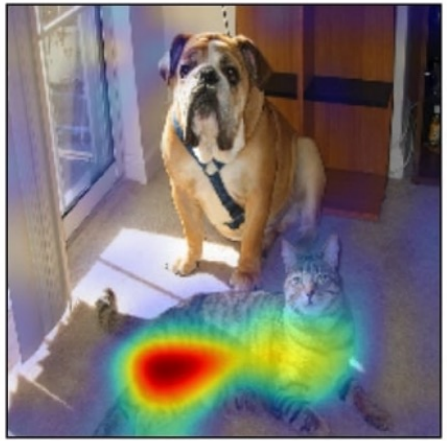
Given an input $x \in \mathbb{R}^d$ with corresponding output $y \in \mathbb{R}$, and a target model $f \in \mathcal{F}$: $\mathbb{R}^d \to \mathbb{R}$.

The explanation model g: $\mathcal{F} \times \mathbb{R}^d \to \mathbb{R}^d$ provides importance scores $g(f, x)$ for each input features.

x

$f(x)$: cat

$f(x)$: *VGG-16*



$g(f, x)$

Preliminaries – SimpleGrad

[In the preliminary, we use the superscript e to denote a specific sample to explain.]

One choice for the explanation model is the partial derivative of $f(x)$ with respect to x:

$$g^{sim}(f, x^e) \coloneqq \frac{\partial f(x)}{\partial x}\Big|_{x=x^e}$$

Sometimes, explanations take the form of element-wise product of inputs and the gradients:

$$g^{inp}(f, x^e) \coloneqq x^e \; \frac{\partial f(x)}{\partial x}\Big|_{x=x^e}$$

Preliminaries – SmoothGrad[1]

SmoothGrad (SG) alleviate the impacts of noise:

1) Take random samples around the input x^e

$$x^e + \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, \sigma^2)$$

2) take the average of the resulting heatmaps.

$$g^{SG}(f, x^e) = \frac{1}{K} \sum_{k=1}^{K} g^{sim}(f, x^e + \epsilon_k)$$

$$g^{sim}(f, x^e) := \frac{\partial f(x)}{\partial x}\Big|_{x=x^e}$$

$$g^{SG}(f, x^e) = \left[\int_z k(x^e, z)dz\right]^{-1} \int_z g(f, z)k(x^e, z)dz$$

[1] Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." *2017*.

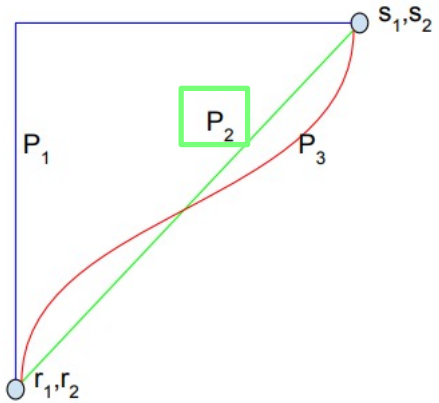
Preliminaries – IntegratedGrad[2]

Integrate the gradient along the path from the baseline to the input.

$$g^{IG}(f, x^e) \coloneqq (x^e - x_0) \times \int_{t=0}^{1} g^{sim}\left(f, x_0 + t(x^e - x_0)\right) dt$$

$t = 0$: $x_0 + t(x^e - x_0) = x_0$

$t = 1$: $x_0 + t(x^e - x_0) = x^e$



[2] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." 2017.

Preliminaries – Completeness[3]

One desiderata of explanation - **Completeness**

$$\sum_{i=1}^{d} g(f,x)_i = f(x) - f(x_0)$$

Methods satisfying completeness:

IntegratedGrad, LRP, Shapley Values, …

Methods unsatisfying completeness:

SimpleGrad, SmoothGrad, …

[3] Ancona, Marco, et al. "Towards better understanding of gradient-based attribution methods for deep neural networks." 2017.

What does this paper do?

Infidelity

- Definition of infidelity
- Find the Minimum of the infidelity
- Relate to existing works and propose other alternatives for perturbations

Sensitivity

- Definition of sensitivity
- Relation between sensitivity and infidelity

Infidelity

One desiderata of explanation - **Completeness**[3]

$$\sum_{i=1}^{d} g(f,x)_i = f(x) - f(x_0)$$

``More rigorously'' (sensitivity-n):

$$\sum_{i \in S} g(f,x)_i = f(x) - f(x[x_S = 0])$$

where $x[x_S = a]_j = a\mathbb{I}(j \in S) + x_j\mathbb{I}(j \in S)$

For example:

$$x = [1,2,4,7,9,13];\ S = \{1,3,4\};\ a = 0;$$

$$x[x_S = a] = [0,2,0,0,9,13]$$

[3] Ancona, Marco, et al. "Towards better understanding of gradient-based attribution methods for deep neural networks." 2017.

Infidelity

Completeness:

$$\sum_{i \in S} g(f, x)_i = f(x) - f(x[x_S = 0])$$

Discrepancy of completeness:

$$corr\left(\sum_{i \in S_k} g(f, x)_i\,, \qquad f(x) - f\left(x\left[x_{S_k} = 0\right]\right)\right)$$

Problems:

1. the “default” value is fixed => more general perturbations?

2. $Corr()$ is Pearson Correlation Coefficient. It is hard to optimize (intractable).

Infidelity

Discrepancy of completeness in [3]:

$$corr\left(\sum_{i\in S} g(f,x)_i\,,\qquad [f(x) - f(x[x_S = 0])]\right)$$

Definition of **Infidelity**:

$$INFD(g,f,x) = \mathbb{E}_{I\sim\mu_I}\left[\left(I^T g(f,x) - \left(f(x) - f(x-I)\right)\right)^2\right]$$

$I \in \mathbb{R}^d$ is the perturbation around x.

1. replace fixed perturbations with random variable I.
2. replace correlation with expected mean square error.

Infidelity

The optimal $g^* = argmin_g \, \mathbb{E}_{I\sim\mu_I}\left[\left(I^T g(f,x) - \left(f(x) - f(x-I)\right)\right)^2\right]$

If I satisfies that $\int II^T d\mu_I$ is invertible:

$$g^*(f,x) = \left(\int II^T d\mu_I\right)^{-1}\left(\int II^T IG(f,x,I) d\mu_I\right)$$

$IG(f,x,I) = \int_{t=0}^{1} \nabla f(x + (t-1)I) dt$ is Integrated Gradient.

Recall: the SmoothedGard:

$$g^k(f,x) = \left[\int_z k(x,z)\, dz\right]^{-1} \int_z g(f,z)k(x,z) dz$$

$g^*(f,x)$ can be considered as applying SmoothGrad on Integrated Gradients, where kernel is not Gaussian kernel but II^T.

Prove: the optimal $g^* = argmin_g \, \mathbb{E}_{I\sim\mu_I}\left[\left|\left|I^T g(f,x) - \left(f(x) - f(x-I)\right)\right|\right|^2\right]$

$$= argmin_g \int \left|\left|I^T g(f,x) - \left(f(x) - f(x-I)\right)\right|\right|^2 d\mu_I$$

$$f(x) - f(x-I) = \int_{x-I}^{x} \nabla_x f(u)du = \int_0^I \nabla_x f(x-I+u)du \overset{t=\frac{u}{I}}{\Longrightarrow} I^T \int_0^1 \nabla_x f(x-I+tI)dt$$

$$= argmin_g \int \left|\left|I^T g(f,x) - I^T \int_0^1 \nabla_x f(x-I+tI)dt\right|\right|^2 d\mu_I$$

To set the first order derivative to 0, and denote $IG(f,x,I) = \int_0^1 \nabla_x f(x-I+tI)dt$

$$2\int II^T\left(g^*(f,x) - IG(f,x,I)\right)d\mu_I = 0$$

$$\int II^T g^*(f,x)d\mu_I = \int II^T IG(f,x,I)d\mu_I$$

$$g^*(f,x) = \left(\int II^T d\mu_I\right)^{-1}\left(\int II^T IG(f,x,I)\right)d\mu_I$$

Infidelity

Definition of Infidelity:

$$INFD(g, f, x) = \mathbb{E}_{I \sim \mu_I}\left[\left(I^T g(f, x) - \left(f(x) - f(x - I)\right)\right)^2\right]$$

Potential choices of I:

1. Difference to baseline(s): (x_0 can be random variable)

$$I = x - x_0$$

2. Subset of difference to baseline: for fixed subset $S \subseteq [d]$

$$I_S = x - x[x_S = (x_0)_S]$$

3. Difference to noisy baseline: ($\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a zero mean random vector)

$$I = x - (x_0 + \epsilon)$$

Infidelity

$$INFD(g, f, x) = \mathbb{E}_{I \sim \mu_I}\left[\left(I^T g(f, x) - \left(f(x) - f(x - I)\right)\right)^2\right]$$

Varying I to recover existing works:

1. If $I = x - x_0$ is deterministic:

   $g^*(f, x) \odot I$ satisfies completeness, can be IG, LRP and DeepLIFT.

2. If $I_\epsilon = \epsilon \cdot e_i$, where e_i is a coordinate basis vector:

   $\lim_{\epsilon \to 0} g^*_{\epsilon_i}(f, x) = \nabla_x f(x)_i$ is the gradient explanation.

3. If $I = e_i \odot x$:

   $g^*(f, x) \odot x$ is the occlusion-1 explanation.

[occlusion-1 explanation replaces one feature x_i at the time with a zero baseline and measuring the effect of this perturbation on the target output.]

Infidelity

$$INFD(g, f, x) = \mathbb{E}_{I \sim \mu_I}\left[\left(I^T g(f, x) - \left(f(x) - f(x - I)\right)\right)^2\right]$$

[Key: set infidelity to 0.]

P2. If $I_{\epsilon_i} = \epsilon e_i$, where e_i is a coordinate basis vector:

$\epsilon g^*(f, x)_i = f(x) - f(x - \epsilon e_i)$.

As $\epsilon \to 0$: $\lim_{\epsilon \to 0} g^*_{\epsilon_i}(f, x) = \lim_{\epsilon \to 0} \frac{f(x) - f(x - \epsilon e_i)}{\epsilon} = \nabla_x f(x)_i$ is the gradient explanation along the i-th coordinate.

P3. If $I = x \odot e_i$:

$x_i g^*(f, x)_i = f(x) - f(x | x_i = 0)$ is the occlusion-1 explanation.

Infidelity

$$INFD(g, f, x) = \mathbb{E}_{I\sim\mu_I}\left[\left(I^T g(f, x) - \left(f(x) - f(x - I)\right)\right)^2\right]$$

Varying I to recover existing works:

4. If $I = h_x(Z)$, where $Z \in \{0,1\}^d$. When $P(Z = z) \propto \frac{d-1}{\binom{d}{||z||_1}||z||_1\left(d-||z||_1\right)}$

$g^*(f, x)\odot x$ is the Shapley value.

$h_x(Z)$: $\{0,1\}^d \to \mathbb{R}^d$ [selects subset], assume $x = [0.5, 20, 7, \dots, 13]$

$$z_1 = [1,1,1, \dots, 1], \qquad h_x(z_1) = x = [0.5, 20, 7, \dots, 13]$$

$$z_2 = [0,0,0, \dots, 0], \qquad h_x(z_2) = 0 = [0,0,0, \dots, 0]$$

$$z_3 = [1,0,0, \dots, 1], \qquad h_x(z_3) = [0.5, 0, 0, \dots, 13]$$

Infidelity

$$INFD(g, f, x) = \mathbb{E}_{I \sim \mu_I}\left[\left(I^T g(f, x) - \left(f(x) - f(x - I)\right)\right)^2\right]$$

Varying I for new possible explanations. (Used in experiments)

[Noisy Baseline]

Set the baseline to be a Gaussian random vector centered around a certain baseline.

$$I = x - (x_0 + \epsilon)$$

[Square Removal] (image only)

$I = h_x(Z)$ where the perturbation Z has a uniform distribution over square patches.

$$I = h_x(Z), Z \sim Uniform$$

Sensitivity

(Vector) For the j-th coordinate, the sensitivity of explanation is defined by the gradient:

$$[\nabla_x g(f,x)]_j = \lim_{\epsilon \to 0} \frac{g(f, x + \epsilon e_j) - g(f,x)}{\epsilon}$$

(Scalar) Compute the norm of the gradient:

$$SENS_{grad}(g,f,x,r) = \sup_{||\delta|| \leq r} ||\nabla_x g(f, x+\delta)||$$

Related to local Lipschitz continuity:

$$SENS_{lips}(g,f,x,r) = \sup_{||\delta|| \leq r} \frac{||g(f,x) - g(f,x+\delta)||}{||\delta||}$$

[If an explanation has locally uniformly bounded gradients, it is locally Lipshitz continuous as well.]

Sensitivity

In this paper, max-sensitivity is proposed:

$$SENS_{max}(g, f, x, r) = \max_{||\delta|| \le r} ||g(f, x+\delta) - g(f, x)||$$

$$\max_{||\delta|| \le r} ||g(f, x+\delta) - g(f, x)|| \le \sup_{||\delta|| \le r} \frac{||g(f, x) - g(f, x+\delta)||}{||\delta||} \cdot r = SENS_{lips}(g, f, x, r) \cdot r$$

Local Lipschitz continuity can be unbounded when using ReLU, but max-sensitivity is always finite.

[Can be estimated by Monte-Carlo sampling in experiments.]

Sensitivity

The max-sensitivity is defined as:

$$SENS_{max}(g, f, x, r) = \max_{||\delta|| \leq r} \left| ||g(f, x + \delta) - g(f, x)|| \right|$$

Remarks:

1. Sensitivity is only one of desiderata.
2. Sensitivity is somehow “nature” of the target models and explanations.

It is nonsense to minimize the max-sensitivity only. We need to consider the fidelity and sensitivity at the same time.

Relation between Sensitivity and Infidelity

The smoothed explanation has less sensitivity and infidelity:

$$g^k(f,x) = \int_z g(f,z)k(x,z)dz$$

$g^k(f,x)$ is less sensitive than the original sensitivity:

$$SENS_{max}\left(g^k,f,x,r\right) \leq \int_z SENS_{max}(g,f,x,r)\,k(x,z)dz$$

$g^k(f,x)$ is less infidelity than the original infidelity (when $\frac{C_2}{1-2\sqrt{C_1}} \leq 1$):

$$INFD\left(g^k,f,x\right) \leq \frac{C_2}{1-2\sqrt{C_1}} \int_z INFD(g,f,z)k(k,z)dz$$

$$C_1 = \max_x \frac{\int_I \int_z (f(z) - f(z-I) - [f(x) - f(x-I)])^2 k(x,z)dz d_{\mu_I}}{\int_I \int_z (I^T g(f,z) - [f(x) - f(x-I)])^2 k(x,z)dz d_{\mu_I}}$$

$$C_1 = \max_x \frac{\int_I \left(\int_z \{I^T g(f,z) - [f(x) - f(x-I)]\} k(x,z)dz\right)^2 d_{\mu_I}}{\int_I \int_z (I^T g(f,z) - [f(x) - f(x-I)])^2 k(x,z)dz d_{\mu_I}}$$

Relation between Sensitivity and Infidelity

To prove $SENS_{max}(g^k, f, x, r) \leq \int_z SENS_{max}(g, f, x, r)\, k(x, z)dz$

$$SENS_{max}(g^k, f, x, r) = \max_{||\delta|| \leq r} \left|\left| g^k(f, x+\delta) - g^k(f, x) \right|\right|$$

$$= \max_{||\delta|| \leq r} \left|\left| \int_z [g(f, z+\delta) - g(f, z)] k(x, z)dz \right|\right|$$

$$\leq \max_{||\delta|| \leq r} \int_z ||g(f, z+\delta) - g(f, z)|| k(x, z)dz$$

$$\leq \int_z \max_{||\delta|| \leq r} \left[\left| |g(f, z+\delta) - g(f, z)| \right|\right] k(x, z)dz$$

$$= \int_z SENS_{max}(g, f, z, r) k(x, z)dz$$

$\phi(x) = ||x||$ is convex

$$\phi\left(\int_x h(x)dx\right) \leq \int_x \phi(h(x))dx$$

Experiments

Dataset: MNIST, CIFAR-10, ImageNet

Explanation methods: Grad, IG, GBP, SHAP, and –SG version

Datasets	MNIST	
Methods	$SENS_{MAX}$	INFD
Grad	0.86	4.12
Grad-SG	0.23	1.84
IG	0.77	2.75
IG-SG	0.22	1.52
GBP	0.85	4.13
GBP-SG	0.23	1.84
Noisy Baseline	0.35	0.51

(a) Results for local explanations on MNIST dataset.

Datasets	MNIST		Cifar-10		Imagenet	
Methods	$SENS_{MAX}$	INFD	$SENS_{MAX}$	INFD	$SENS_{MAX}$	INFD
Grad	0.56	2.38	1.15	15.99	1.16	0.25
Grad-SG	0.28	1.89	1.15	13.94	0.59	0.24
IG	0.47	1.88	1.08	16.03	0.93	0.24
IG-SG	0.26	1.72	0.90	15.90	0.48	0.23
GBP	0.58	2.38	1.18	15.99	1.09	0.15
GBP-SG	0.29	1.88	1.15	13.93	0.41	0.15
SHAP	0.35	1.20	0.93	5.78	–	–
Square	0.24	0.46	0.99	2.27	1.33	0.04

(b) Results for global explanations on MNIST, Cifar-10 and imagenet.

Table 1: Sensitivity and Infidelity for local and global explanations.

Experiments

Qualitative experiments