

Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution

Gary S. W. Goh¹ Sebastian Lapuschkin² Leander Weber²
Wojciech Samek² Alexander Binder¹

¹ISTD Pillar
Singapore University of Technology and Design

²Fraunhofer Heinrich Hertz Institute

ICPR2020, 12 January 2021



Introduction

Motivations:

- difficult to explain for deep neural network's decisions due to black-box behavior
- poor input-to-output inference and interpretability
- lack of trust between humans and AI systems

Attributions: measure the contribution of the models' output explained in terms of the input variables. For e.g. image classification

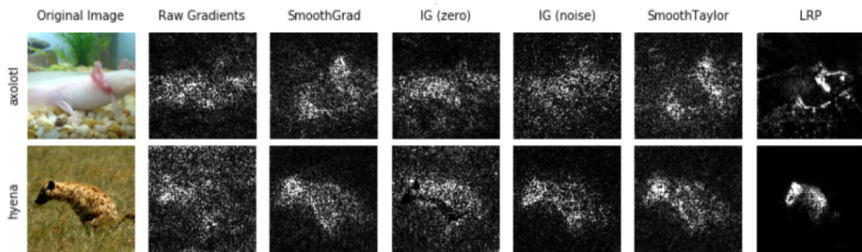


Figure 1: Example of saliency maps from different attribution methods

Preliminaries: Integrated Gradients (IG)

Given a deep neural network represented by a function f for input x :

Integrated Gradients (Sundararajan et al. 2017)

$$IG_i(x, z) := (x_i - z_i) \times \int_{\alpha=0}^1 \frac{\partial f(z + \alpha \times (x - z))}{\partial x_i} d\alpha \quad (1)$$

$$\approx (x_i - z_i) \times \frac{1}{M} \sum_{m=1}^M \frac{\partial f(z + \frac{m}{M} \times (x - z))}{\partial x_i} \quad (2)$$

where $\frac{\partial f(x)}{\partial x_i}$ is the gradient of f in the i^{th} dimension, and z is a selected input baseline.

Satisfies two key axioms:

- **implementation invariance**: independent on model's structure
- **completeness**: attributions add up to the output difference between input x and baseline z (i.e. $\sum_i IG_i(x, z) = f(x) - f(z)$)

Preliminaries: Integrated Gradients Baselines

Question: How to choose baseline z ?

- zero vector (absence of input features)
 - uniform noise¹
- } statistical outliers!

e.g. image baseline inputs:



Figure 2: Black image



Figure 3: Uniform noise

Integrated Gradients with Uniform Noise Baseline

To address the issue of which random noise to be chosen, we take the average of multiple attributions using N different random noise:

$$\overline{IG}_{noise}(x) = \frac{1}{N} \sum_{n=1}^N IG(x, z^{(n)}) \quad (3)$$

¹<https://github.com/ankurtalv/Integrated-Gradients/>

Preliminaries: SmoothGrad

A technique which compute an attribution map by averaging over multiple attributions maps of an arbitrary attribution method (denoted as \mathcal{M}) with multiple N' noised inputs, creating visually sharper attribution maps:

SmoothGrad (Smilkov et al. 2017)

$$\text{SmoothGrad}(x) = \frac{1}{N'} \sum_{n=1}^{N'} \mathcal{M}(x + \epsilon), \quad (4)$$

where $\epsilon \sim \mathcal{N}(0, \sigma'^2)$

Observation: Gaussian noise parameter σ'^2 needs to be carefully selected to get best results

SmoothTaylor Definition

Given a deep neural network represented by a function f for input x :

SmoothTaylor

$$\text{SmoothTaylor}_i(x) := \int_{z \in S} (x_i - z_i) \frac{\partial f(z)}{\partial x_i} dz \quad (5)$$

$$\approx \frac{1}{R} \sum_{r=1}^R (x_i - z_i^{(r)}) \frac{\partial f(z^{(r)})}{\partial x_i} \quad (6)$$

where $z^{(r)} \sim S$

and $z \in S$ is a measurable set of selected roots

Two salient differences from IG:

- explanation point z_i is inner product $(x_i - z_i)$ is part of the integral
- integration set S is not a path

Smooth Taylor Derivation

Any arbitrary differentiable function f can be approximated by Taylor's theorem with just the first order term:

Taylor's theorem

$$f(x) \approx f(z) + \sum_i (x_i - z_i) \frac{\partial f(z)}{\partial x_i} \quad (7)$$

This explains how $f(\cdot)$ in point x is different from the output of the same model in point z . Notably, it is an explanation for x relative to z .

Question: How to choose z ?

$$f(x) \approx \frac{1}{R} \sum_{r=1}^R \left[f(z^{(r)}) + \sum_i (x_i - z_i^{(r)}) \frac{\partial f(z^{(r)})}{\partial x_i} \right] \quad (8)$$

We draw several R roots $z^{(r)}$ and take the average.

SmoothTaylor Roots Generation

Inspired by *SmoothGrad*, a simple approach is to inject a random variable ϵ to input x , where ϵ can be drawn from a Gaussian distribution with standard deviation σ being the noise scaling factor

$$z^{(r)} = x + \epsilon \quad (9)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Theorem

If the roots in *SmoothTaylor* are chosen as per Equation (9), then the discrete version of *SmoothTaylor* as given in Equation (6) is a special case of *SmoothGrad* with $\mathcal{M} = \nabla f(x + \epsilon) \cdot \epsilon$.

- z must be chosen not to close or too far from x (carefully balanced)
- SmoothTaylor does not require a selected baseline z vs. IG
- theorem establishes SmoothTaylor as a theoretical bridge between IG and SmoothGrad

Experiment Setup

Goal: Compare empirical performance of SmoothTaylor vs. IG

Scope: Image classification task using ILSVRC2012 ImageNet dataset (first 1000 images of validation set)

- 1000 multi-class image classification
- image pre-processed to be 224×224 pixels

Pre-trained models: DenseNet121 and ResNet152

Hyperparameters:

- IG (zero): $M = 50$
- IG (noise): $M = 50$, and $N = 1, 5, 10, 20$
- SmoothTaylor: $R = 100, 150, 200$, and $\sigma = 3e-1, 5e-1, 7e-1$

Evaluation Metric: Perturbation scores (Sensitivity)

(Samek et al. 2017)

- find ordered sequence $\mathcal{O} = (r_1, r_2, \dots, r_L)$ the top- L most salient non-overlapping regions of kernel size $k \times k$
- apply perturbation function g on most relevant region first (following \mathcal{O}) iteratively L times:

$$\forall 1 \leq l \leq L : x^{(l)} = g(x^{(l-1)}, r_l) \quad (10)$$

- at each step l , we consider P different random perturbation samples and compute the mean score $\bar{y}^{(l)}$:

$$\bar{y}^{(l)} = \frac{1}{P} \sum_{p=1}^P f(x^{(l-1)(p)}) \quad (11)$$

- overall measure: area under perturbation curve (AUPC)

Hyperparameters: $k = 15$, $L = 30$, $P = 50$

Evaluation Metric: Smoothness (ATV)

- apply min-max normalization (absolute values $> 99^{\text{th}}$ percentile clipped off) on attribution map to construct saliency map \mathcal{S}
- given \mathcal{S} as vector of size $h \times w$ to represent every pixel, the Average Total Variation (ATV) of \mathcal{S} is computed as follows:

$$ATV(\mathcal{S}) = \frac{1}{h \times w} \sum_{i,j \in \mathcal{N}} \|\mathcal{S}_i - \mathcal{S}_j\|_p \quad (12)$$

Here, \mathcal{N} defines the set of pixel neighbourhoods (adjacent horizontal and vertical pixels) and $\|\cdot\|$ is the ℓ_p norm. (we use ℓ_1)

- construct Gaussian pyramids (Burt & Adelson, 1983) on \mathcal{S} and the ATV at every steps are called multi-scaled ATVs
- overall measure: area under multi-scaled ATV curve (AUTVC)

Hyperparameters: downscale factor = 1.5,
and minimum pyramid size = 30×30 (total 5 steps)

Evaluation Metrics Curves

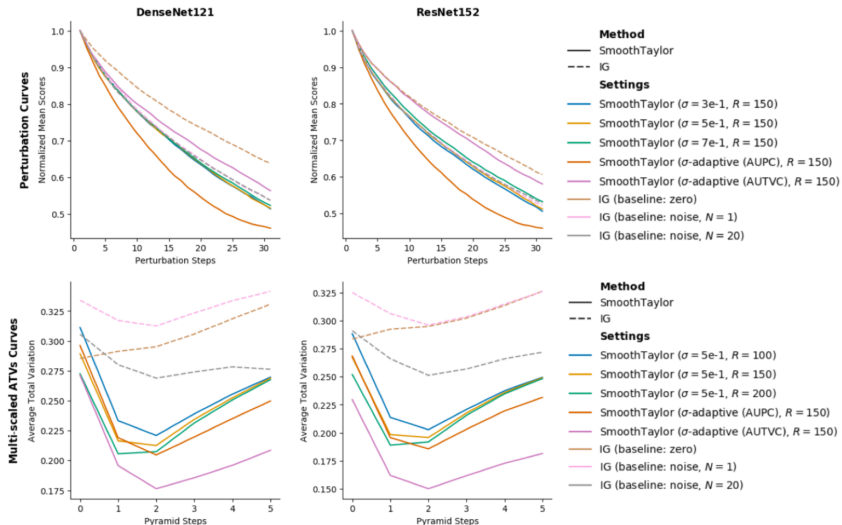


Figure 4: Evaluation metrics curves; the lower the curve the better.

Examples: Saliency Maps

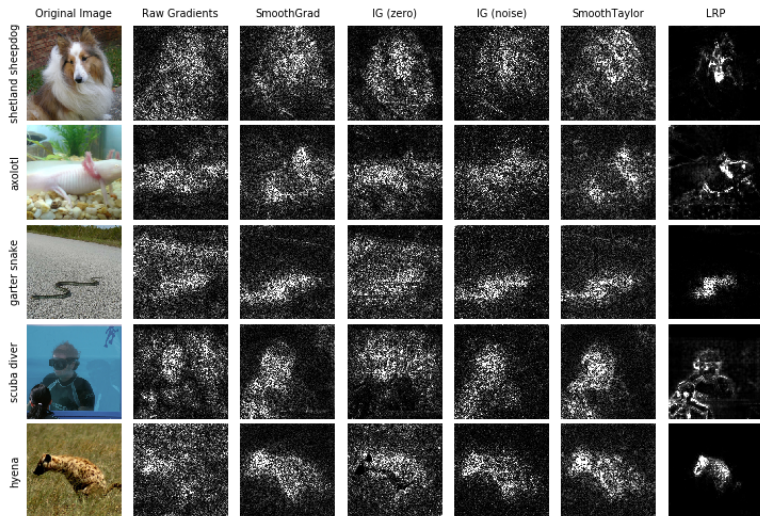


Figure 5: Examples of some saliency maps (visualized attribution maps)

Experiment Results

TABLE I
AREA UNDER THE CURVES RESULTS.
NOTE: LOWER AUPC AND AUTVC IS BETTER.

Attribution Method		Image Classifier Model			
		DenseNet121		ResNet152	
IG					
baseline	N	AUPC	AUTVC	AUPC	AUTVC
zero	-	23.63	1.52	22.87	1.51
noise	1	21.51	1.62	21.05	1.54
	5	21.54	1.52	20.99	1.43
	10	21.46	1.45	21.02	1.37
	20	21.43	1.39	21.02	1.32
SmoothTaylor					
σ		DenseNet121		ResNet152	
		AUPC	AUTVC	AUPC	AUTVC
$3e-1$	100	21.24	1.28	20.83	1.20
	150	21.19	1.24	20.79	1.16
	200	21.13	1.22	20.78	1.14
$5e-1$	100	21.25	1.23	21.00	1.14
	150	21.20	1.19	20.95	1.10
	200	21.13	1.16	20.86	1.07
$7e-1$	100	21.39	1.20	21.37	1.08
	150	21.30	1.15	21.32	1.04
	200	21.30	1.12	21.14	1.01

- IG with noise baseline with large N have huge improvements over IG with zero baseline, but still a little worse as compared to SmoothTaylor
- initial choice of σ values has little effect on performance (we investigate further)
- SmoothTaylor performance improves as R increase due to greater "smoothing" effect

Sensitivity Analysis

TABLE II

AREA UNDER THE CURVES RESULTS FOR *SmoothTaylor* WITH EXTREME HYPERPARAMETER VALUES.

NOTE: LOWER AUPC AND AUTVC IS BETTER.

SmoothTaylor		Image Classifier Model			
Hyperparameters		DenseNet121		ResNet152	
σ	R	AUPC	AUTVC	AUPC	AUTVC
5e-1	10	21.74	1.55	21.43	1.43
1e-4	100	23.45	1.79	23.00	1.55
1e-3	100	23.60	1.53	23.14	1.48
1e-2	100	23.90	1.57	23.46	1.23
1e-1	100	22.03	1.43	21.44	1.22
1	100	21.88	1.17	22.16	1.04
2	100	23.54	1.19	24.48	1.27

- experiment with σ to be as high as $2e+0$ and as low as $1e-4$, while fixing R to be 100
- when σ is too small ($< 1e-3$) or big ($2e+0$), AUPC is worse
- can be explained from gradient shattering effects (Balduzzi et al. 2017) across multiple linearity zones
- optimal window range of is sample-dependent; support the claim that *sigma* needs to be carefully calibrated

Adaptive Noising

Algorithm 1: Adaptive Noising

Parameters: Max. iterations i_{max} , learning rate α , learning decay γ , max. stop count s_{max}

Input : Input x , root size R , model f

Output : Optimal σ^* value

begin

```
 $\sigma \leftarrow \frac{1}{\sqrt{N}} \sum |x|;$ 
 $AUC \leftarrow \text{ComputeAUC}(x, R, f, \sigma);$ 
 $i \leftarrow 1; s \leftarrow 0; \sigma^* \leftarrow \sigma; AUC^* \leftarrow AUC;$ 
while  $i \leq i_{max}$  do
   $AUC_s \leftarrow \text{ComputeAUC}(x, R, f, |\sigma + \alpha|);$ 
  if  $AUC_s > AUC$  then
     $\sigma \leftarrow |\sigma - \alpha|;$ 
     $AUC_s \leftarrow \text{ComputeAUC}(x, R, f, \sigma);$ 
  else
     $\sigma \leftarrow |\sigma + \alpha|;$ 
  end
  if  $AUC_s > AUC$  then
    if  $s \leq s_{max}$  then
       $\alpha \leftarrow \alpha * \gamma; s \leftarrow s + 1;$ 
    else
      break
    end
  else
     $s \leftarrow 0;$ 
    if  $AUC_s < AUC^*$  then
       $AUC^* \leftarrow AUC_s; \sigma^* \leftarrow \sigma;$ 
    end
  end
   $AUC \leftarrow AUC_s; i \leftarrow i + 1;$ 
end
end
```

- iterative heuristic line search
- find σ^* so as to minimize for either AUPC or AUTVC
- SmoothTaylor with adaptive noising achieves best performance

Hyperparameters: $R = 150$,
 $i_{max} = 20$, $s_{max} = 3$, $\alpha = 0.1$,
 $\gamma = 0.9$

TABLE III
AREA UNDER THE CURVES RESULTS WITH ADAPTIVE NOISING.
NOTE: LOWER AUPC AND AUTVC IS BETTER.

SmoothTaylor Hyperparameters	Image Classifier Model				
	DenseNet121	ResNet152			
σ	R	AUPC	AUTVC	AUPC	AUTVC
Adaptive-AUPC	150	19.55	1.14	19.30	1.05
Adaptive-AUTVC	150	22.14	0.99	22.52	0.85

Conclusion

Paper contributions:

- present SmoothTaylor as a theoretical concept bridge between IG and SmoothGrad
- emphasize smoothness as a key quality measure for attribution and introduce multi-scaled ATVs as a new evaluation metric
- empirically show that SmoothTaylor can produce more relevance-sensitive and less noisy attribution maps vs. IG
- further propose adaptive noising as a hyperparameter tuning technique to optimize SmoothTaylor's performance

References

- M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in 34th International Conference on Machine Learning, ICML 2017, vol. 7, Sydney, Australia, 2017, pp. 5109–5118.
- D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg, “SmoothGrad: Removing noise by adding noise,” in Workshop on Visualization for Deep Learning, ICML, 2017.
- W. Samek, A. Binder, G. Montavon, S. Bach, and Klaus-Robert Muller, “Evaluating the Visualization of What a Deep Neural Network Has Learned,” IEEE Transactions on Neural Networks and Learning Systems, vol. 8, no. 11, pp. 2660 – 2673, 2017.
- P. J. Burt and E. H. Adelson, “The Laplacian Pyramid as a Compact Image Code,” IEEE Transactions on Communications, vol. VOL. COM- 3, no. 4, pp. 532–540, 1983.
- D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. Wan-Duo Ma, and B. McWilliams, “The Shattered Gradients Problem: If resnets are the answer, then what is the question?” in Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.

End

Q & A

Thank you for your attention.

Poster session: **PS T1.3** on DAY 2 – January 13, ,2021

Source code: <https://github.com/garygsw/smooth-taylor>