Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution

Gary S. W. Goh[1], Sebastian Lapuschkin[2], Leander Weber[2], Wojciech Samek[2] and Alexander Binder[1]

[1] ISDT Pillar, Singapore University of Technology and Design, [2] Fraunhofer Heinrich Hertz Institute

Abstract

Integrated Gradients as an attribution method for deep neural network models offers simple implementability. However, it suffers from noisiness of explanations which affects the ease of interpretability. The SmoothGrad technique is proposed to solve the noisiness issue and smoothen the attribution maps of any gradient-based attribution method. In this paper, we present SmoothTaylor as a novel theoretical concept bridging Integrated Gradients and SmoothGrad, from the Taylor's theorem perspective. We apply the methods to the image classification problem, using the ILSVRC2012 ImageNet object recognition dataset, and a couple of pretrained image models to generate attribution maps. These attribution maps are empirically evaluated using quantitative measures for sensitivity and noise level. We further propose adaptive noising to optimize for the noise scale hyperparameter value. From our experiments, we find that the SmoothTaylor approach together with adaptive noising is able to generate better quality saliency maps with lesser noise and higher sensitivity to the relevant points in the input space as compared to Integrated Gradients.

Overview

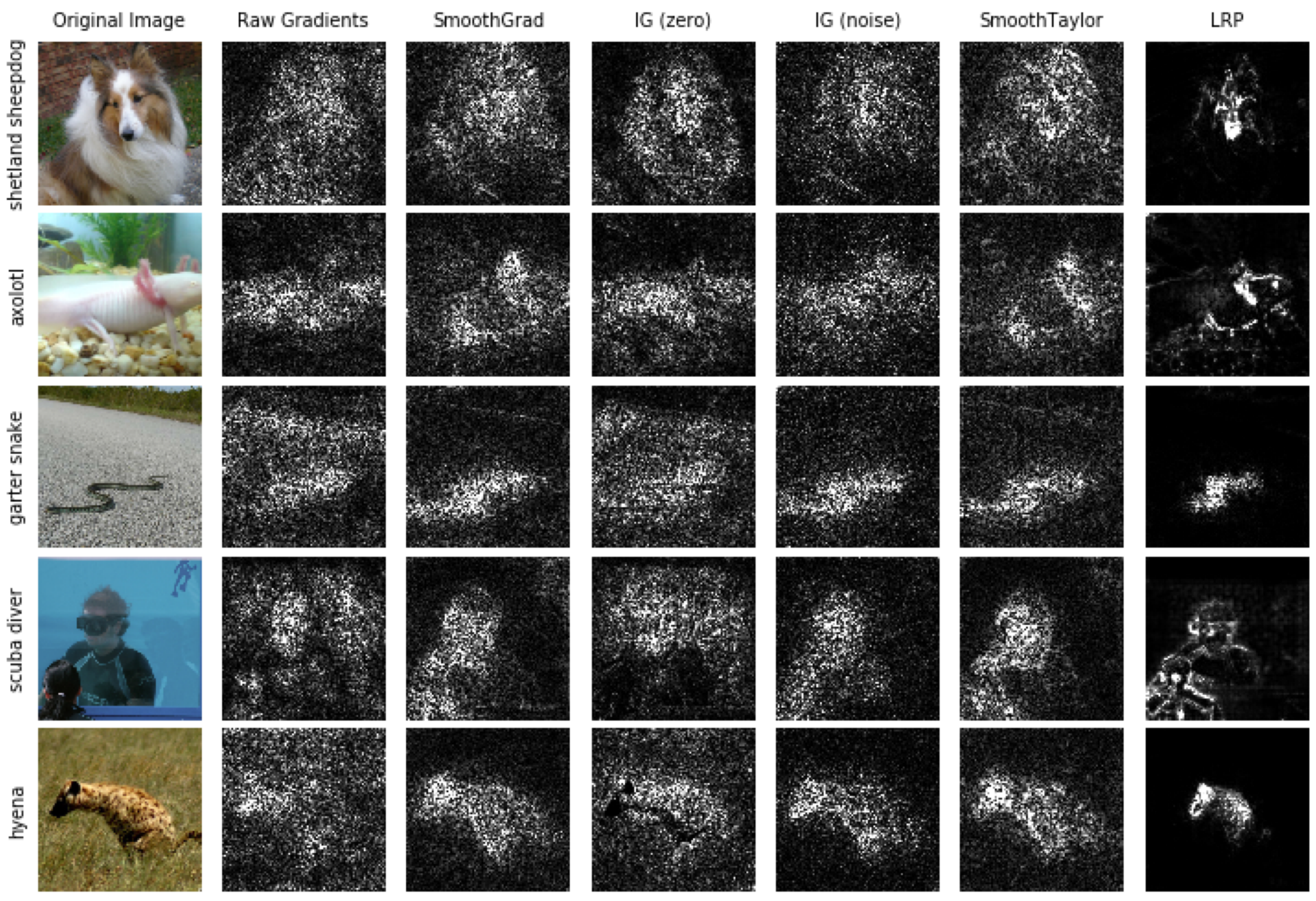
Motivation:

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

Input
e.g. images, text
Deep Neural Network
(black box)
Output
e.g. class, text

Deep Neural Networks Attribution:

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



Preliminaries

Integrated Gradients (IG)

With a deep neural network represented by a function f for input x, the integrated gradients for the i^{th} dimension is defined as (Sundararajan: et al., 2017):

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

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

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:

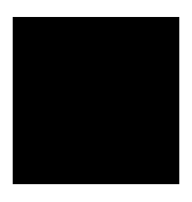
- implementation invariance: independent on model's structure
- completeness: attributions add up to the output difference between input x and baseline z (i.e. $\sum IG_i(x,z) = f(x) - f(z)$)

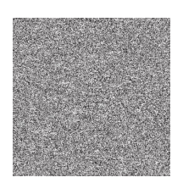
Selection of baseline z:

1. Zero vector

2. Uniform noise

e.g. Images:





IG with uniform noise baselines:

Take average of multiple N IG attribution maps

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

Key observations:

- IG with uniform noise as baseline has higher quality attribution maps
- Selected baselines are usually statistically outliers in the input space, thus making explanations to such points seem irrelevant

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 (Smilkov et al. 2017) :

$$SmoothGrad(x) = \frac{1}{N'}\sum_{n=1}^{N'} \mathcal{M}(x+\epsilon),\ \epsilon \sim \mathcal{N}(0, \sigma'^2) \qquad (4)$$

Observations:

- able generate visually shaper attribution maps
- gaussian noise parameter σ'needs to be carefully selected to get best results

SmoothTaylor

Definition:

With a deep neural network represented by a function f for input x, the SmoothTaylor for the i^{th} dimension is defined as:

$$SmoothTaylor_i(x) := \int_{z\in S} (x_i - z_i)\frac{\partial f(z)}{\partial x_i} dz \qquad (5)$$

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

where $z^{(r)} \sim S$ and $z \in S$ is a measurable set of selected roots

Two salient differences from IG's formulation:

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

Derivation:

Any arbitrary differentiable function f can be approximated by Taylor'stheorem with just the first order term, and can be statistically improved by drawing multiple R roots and take the average to improve the power of the approximation:

$$f(x) \approx f(z) + \sum_i (x_i - z_i)\frac{\partial f(z)}{\partial x_i} \approx \frac{1}{R}\sum_{r=1}^{R}\left[f(z^{(r)}) + \sum_i (x_i - z_i^{(r)})\frac{\partial f(z^{(r)}))}{\partial x_i}\right] \qquad (7)$$

SmoothTaylor is derived from Eq (7) by carefully selecting a set of roots such that $f(x)$ in the LHS can be cancelled out with $f(z^{(r)})$ in the RHS. Inspired by SmoothGrad, we inject a random variable ϵ to input x, where ϵ can be drawn:

$$z^{(r)} = x + \epsilon, \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2) \qquad (8)$$

Theorem: If the roots in SmoothTaylor are chosen as per Eq (8), then the discrete version of SmoothTaylor as given in Eq (6) is a special case of SmoothGrad with $\mathcal{M} = \nabla f(x+\epsilon)\cdot\epsilon$

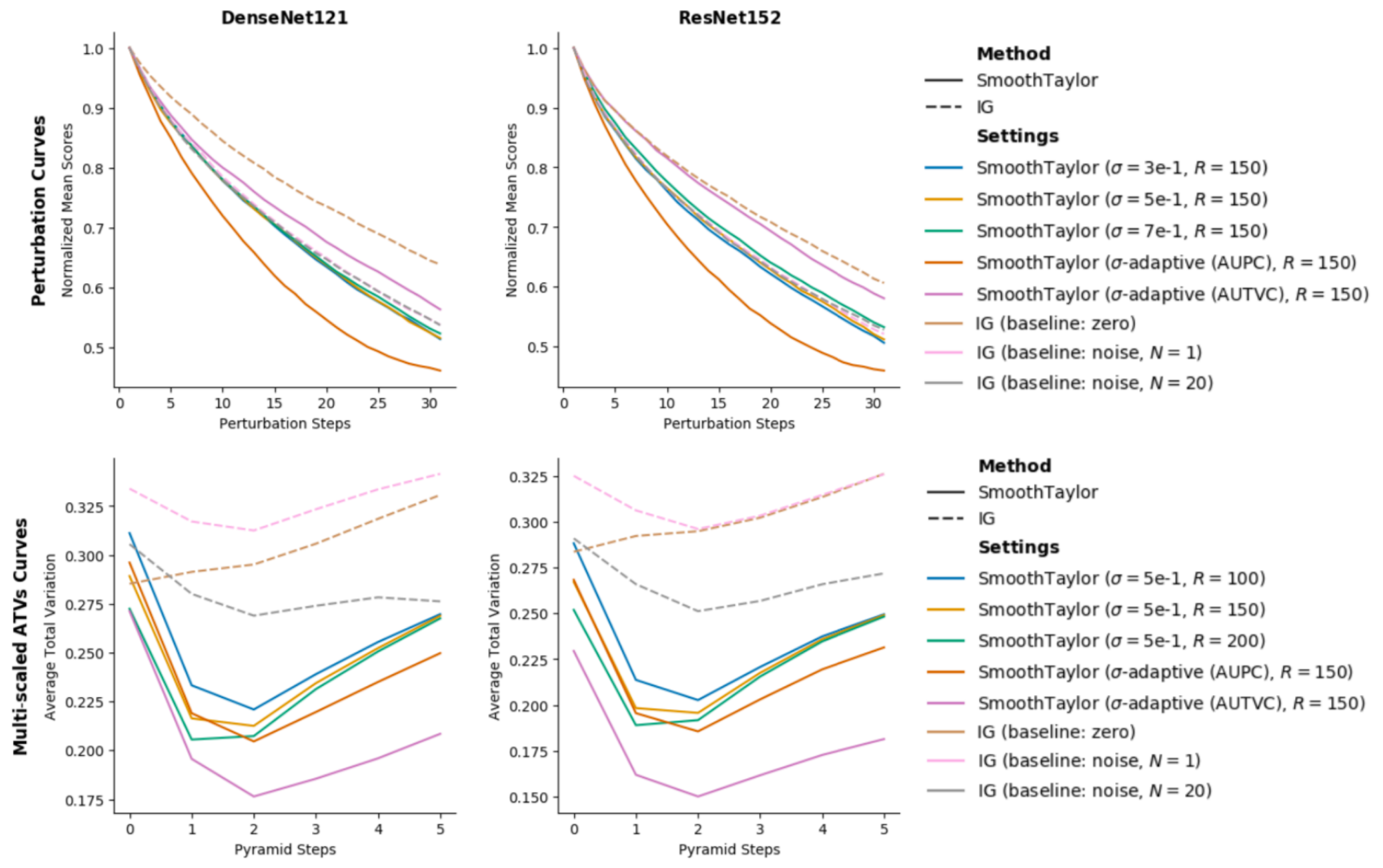
Key contributions:

- SmoothTaylor does not require a selected baseline z as compared to IG
- Theorem establishes SmoothTaylor as a theoretical bridge between IG and SmoothGrad
- Further propose adaptive noising as a hyperparameter tuning technique for σ

Experiments & Results

Setup

- **Scope:** Image classification on using ILSVRC2012 ImageNet dataset with pre-trained image classifiers DenseNet121 and ResNet152
- **Hyperparameters:** IG (zero): M = 50; IG (noise): M = 50, and N = 1,5,10,20; SmoothTaylor: R = 100, 150, 200, and σ = 3e-1, 5e-1, 7e-1
- **Evaluation metrics**: Area Under Perturbation Curves (AUPC) & Area Under Multi-scaled Average Total Variations Curves (AUTVC) **(lower the 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		DenseNet121		ResNet152	
σ	R	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
Adaptive-AUPC	150	**19.55**	1.14	**19.30**	1.05
Adaptive-AUTVC	150	22.14	**0.99**	22.52	**0.85**

Refences:

- 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. Vie ́gas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," in Workshop on Visualization for Deep Learning, ICML, 2017.