



# SmoothGrad: removing noise by adding noise

이상용 / 2020-02-28



Computational Data Science LAB



# SmoothGrad: removing noise by adding noise

Computational Data Science LAB

목차

1. Introduction
2. Gradients as sensitivity maps
3. Experiments



논의사항 및  
결정사항

관련문서

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. ICML Workshop. 2017.



# CONTENTS

1. Introduction
  2. Gradients as sensitivity maps
  3. Experiments
- 
- 

# 01 | Introduction

- Deep neural networks 같은 복잡한 모델을 해석하는 것은 아직까지 어려운 난제이지만 특정 도메인에서 interpretability는 굉장히 중요하게 여겨짐
- classification task에서 “explanation”을 찾는 것은 시스템의 메커니즘을 밝히고 잠재적으로 시스템을 향상시키는 데 도움을 줄 수 있음
- 이미지 분류 문제에서 결과를 해석하기 위해 사용하는 일반적인 방법은 어떤 영역이 결과에 영향을 미쳤는지 확인하는 것 → [sensitivity maps](#), saliency maps, pixel attribution maps 이라고 불림
- 이런 방법은 classification 결과에 영향을 미친 부분을 강조하지만, 동시에 관련이 덜한 noisy한 영역도 강조
- 본 논문은 이러한 noise를 줄이면서 간단하고 다른 sensitivity map 방법에도 적용 가능한 방법을 제안함

## 02 | Gradients as sensitivity maps

- 이미지 분류 문제에서 하나의 이미지 셋에 대한 클래스를 분류하는 문제는 다음과 같이 이해할 수 있음

$$\text{class}(x) = \operatorname{argmax}_{c \in C} S_c(x)$$

$C$  : image set,  $c$  : class,  $S_c$  : class activation function

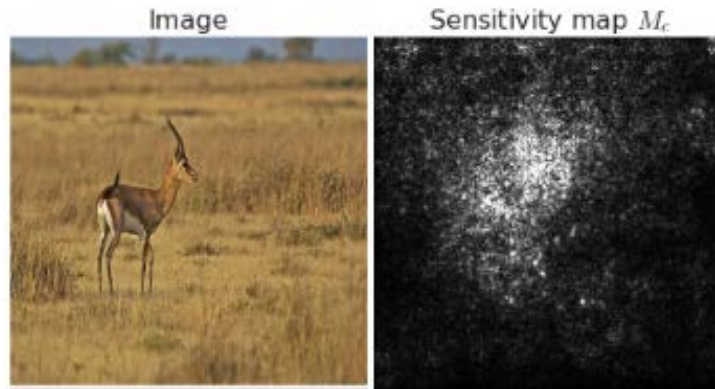
- 만약  $S_c$ 가 piecewise differentiable하다면 image  $x$ 의 sensitivity map  $M_c$ 는 다음과 같이 정의할 수 있음

$$M_c(x) = \partial S_c(x) / \partial x$$

- ✓  $\partial S_c$ 는  $S_c$ 의 gradient를 나타냄
- ✓  $M_c$ 는 sensitivity map
- ✓ 직관적으로,  $M_c$ 는 각 pixel의 작은 변화에 class  $c$ 에 대한 score가 변화하는 정도를 나타냄
- ✓ class  $c$ 에 대해 중요한 pixel은 gradient의 크기가 크고 중요하지 않은 부분은 작을 것이라고 생각할 수 있음

## 02 | Gradients as sensitivity maps

- 실제로,  $M_c$ 는 해당 클래스  $c$ 를 예측하는데 관련이 있는 영역이 부각되어 보임

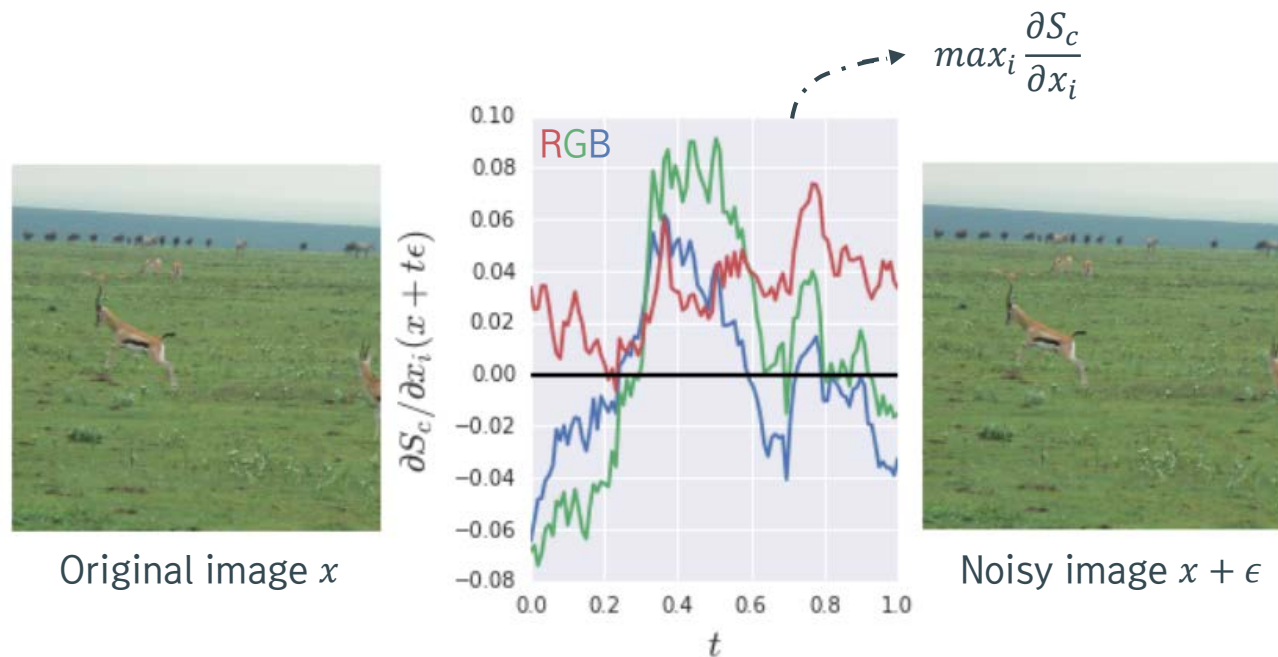


- ✓ 하지만, gradient로 나타낸 sensitivity map은 굉장히 noisy함
- ✓ 본 논문은 이런 노이즈를 본질적으로 의미가 없는 영역이라고 가정하며, 노이즈를 줄이면서 객체를 좀 더 선명하게 부각시킬 수 있는 방법을 제안함

## 02 | Gradients as sensitivity maps

### Smoothing noisy gradients

- 주어진 이미지에 대해서 한 pixel의 gradient fluctuation을 확인

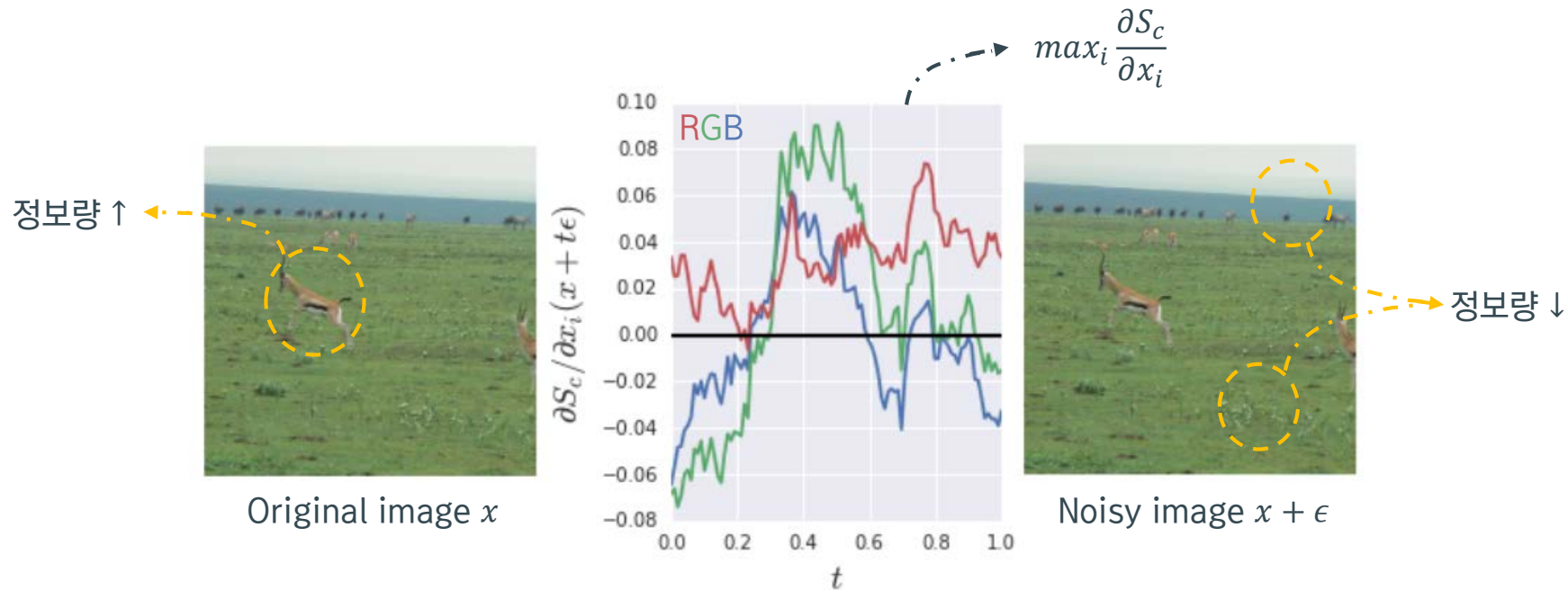


- ✓ 원래의 이미지와 노이즈가 추가된 이미지는 사람이 보기에 큰 차이가 없어 보이며 거의 유사함
- ✓ 노이즈가 추가된 이미지 또한 원래의 이미지와 동일한 클래스(가젤)로 예측 됨

## 02 | Gradients as sensitivity maps

### Smoothing noisy gradients

- 주어진 이미지에 대해서 한 pixel의 gradient fluctuation을 확인



- ✓ 의미있는 영역이라 여겨지는 pixel에 대한 gradient는 강한 fluctuation을 보임
- ✓ Gradient의 변화가 큰 부분 → 정보량이 큼
- ✓ Gradient의 변화가 작은 부분 → 정보량이 적음



## 02 | Gradients as sensitivity maps

### Smoothing noisy gradients

- 새로운 sensitivity map: SMOOTHGRAD 방법을 제안
  - ✓ Gradient  $\partial S_c$ 를 바로 사용하지 않고 Gaussian kernel을 사용하여 gradient의 평균값을 visualization

$$\hat{M}_c(x) = \frac{1}{n} \sum_1^n M_c(x + \mathcal{N}(0, \sigma^2))$$

- ✓  $n$  : 샘플 사이즈
- ✓  $\mathcal{N}(0, \sigma^2)$  : Gaussian noise를 나타냄
- ✓ 핵심 아이디어는 noise가 더해진 유사한 샘플이미지의 평균을 취하여 sensitivity map을 만들
- ✓  $\sigma, n$ 는 SMOOTHGRAD의 하이퍼 파라미터

# 03 | Experiments

## Visualization methods and techniques

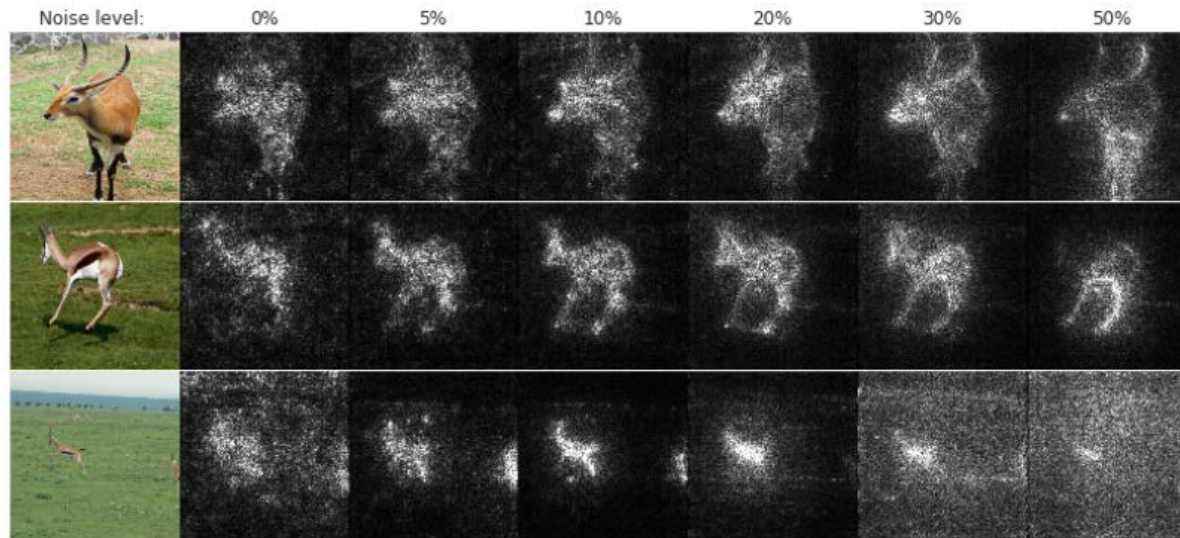
- Absolute value of gradients
  - ✓ Sensitivity map을 만들 때 gradient는 주로 양수, 음수, 절대값이 사용됨
  - ✓ 어떤 값을 사용할 지는 데이터셋의 특징에 달려있음
  - ✓ MNIST같이 같은 색상의 데이터셋은 양수와 음수를 사용하는 것이 적절하며, ImageNet 데이터셋과 같이 여러 색상이 혼재되어 있는 경우 절대값의 사용하는 것이 더 깔끔한 결과를 도출 함
- Capping outlying values Another
  - ✓ Gradient에 대한 다른 속성은 이상치에 대한 것으로, 극단적으로 높은 값을 가지는 이상치가 존재할 때 sensitivity map을 만들면 이상치를 제외한 다른 부분은 어두운 값을 가지게 됨
- Multiplying maps with the input images
  - ✓ 다른 기법은 sensitivity map을 만들 때 gradient value와 실제 pixel value를 곱하는 것

# 03 | Experiments

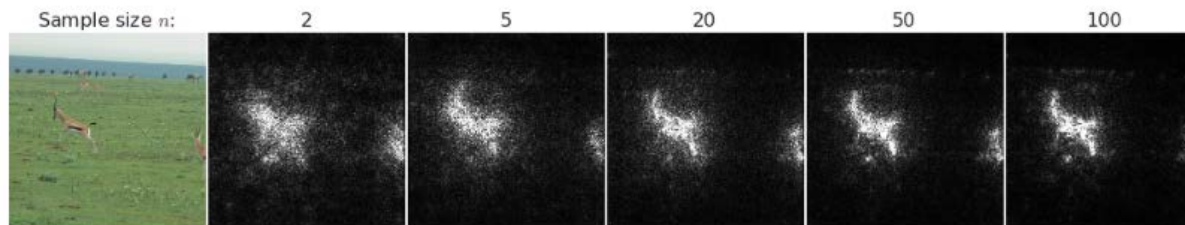
## Effect of noise level and sample size

- Noise  $\sigma$

Noise의 수준 :  $\sigma / (x_{max} - x_{min})$

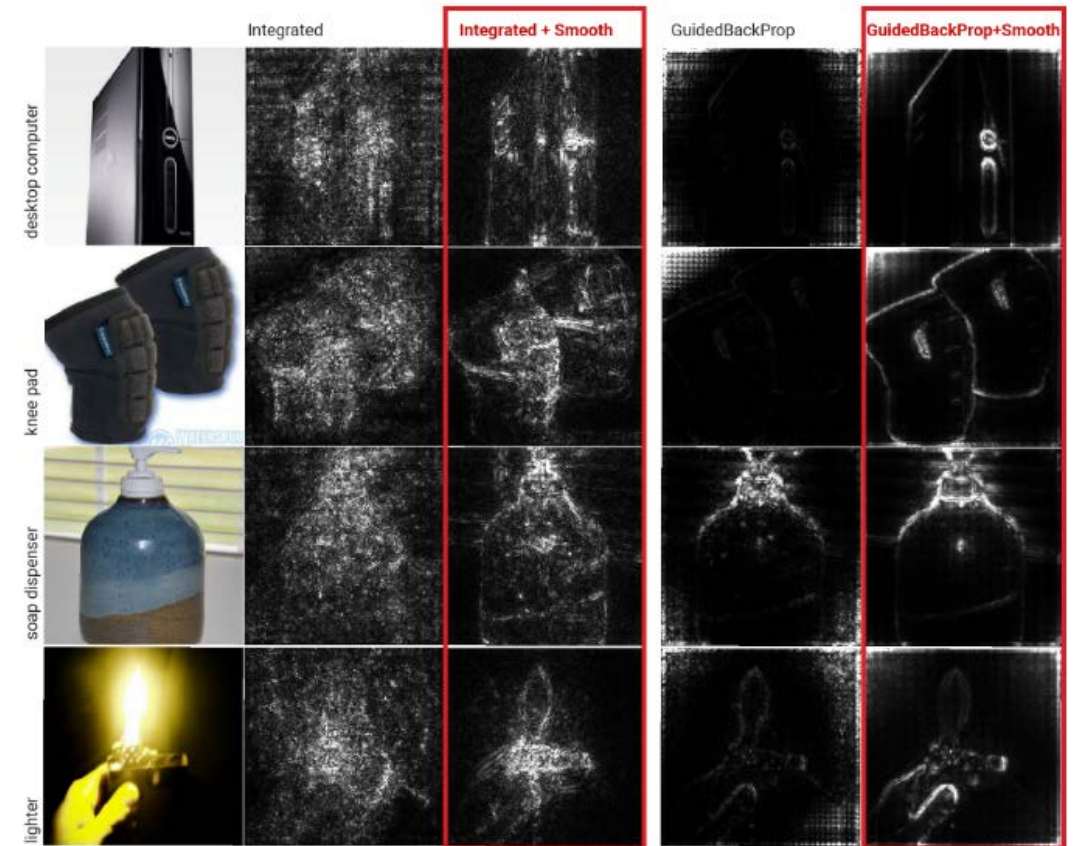
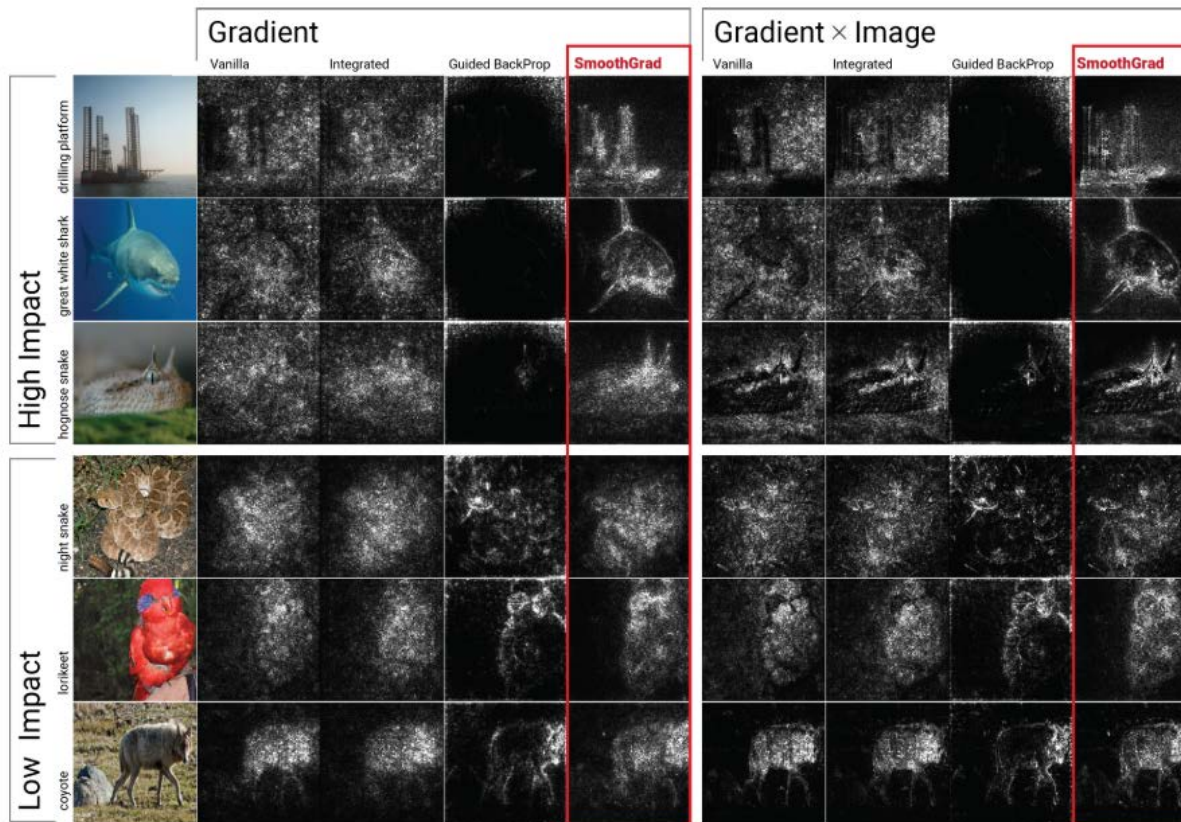


- Sample size  $n$



# 03 | Experiments

- Qualitative comparison to baseline methods
- Combining SmoothGrad with other methods



# Q&A

---

감사합니다.