



Axiomatic Attribution for Deep Networks

이상용 / 2020-03-20



Computational Data Science LAB



Axiomatic Attribution for Deep Networks

Computational Data Science LAB

목차

1. Introduction
2. Two Fundamental Axioms
3. Method: Integrated Gradients
4. Experiments



논의사항 및
결정사항

관련문서

Sundararajan, M., Taly, A., & Yan, Q. (2017, August). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 3319-3328). JMLR. org.



CONTENTS

1. Introduction
 2. Two Fundamental Axioms
 3. Method: Integrated Gradients
 4. Experiments
- 
- 

01 | Introduction

- 딥러닝에서 input feature의 attribution을 구하는 방법들이 많이 연구되고 있음
- 여러 방법들 중에서 크게 두 가지의 기본적인 Axioms를 식별함
 1. *Sensitivity*
 2. *Implementation Invariance*
- Gradient를 사용하는 방법들 중 몇몇은 axioms 중 하나를 위반함
- 본 논문은 두 가지의 Axioms를 위반하지 않는 attribution method인 *Integrated Gradient*를 제안함

02 | Two Fundamental Axioms

- Attribution problem에서 base-line의 필요성을 검토 (ex. DeepLIFT)
 - ✓ Attribution을 시행하는 기준으로 base-line과 output의 결과를 비교
 - ✓ Image에서 base-line은 black image, text에서는 zero embedding vector 등등
- Gradient
 - ✓ Gradient는 DNN에서 모델에 대한 coefficient의 자연스러운 동류이기 때문에 attribution method에 대한 starting point로 보아도 합리적임 → gradient를 기준으로 공리에 초점을 맞춤

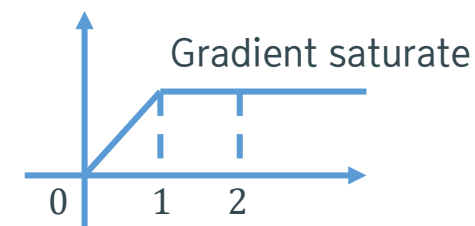
02 | Two Fundamental Axioms

- Axiom: Sensitivity

- ✓ 모든 입력에 대해서 하나의 feature만 달라졌을 때, prediction이 달라진다면 해당 feature의 기여도는 0이 아님
- ✓ 자명한 사실 같아 보이지만 gradient는 이 조건을 만족하지 못함

$$f(x) = 1 - \text{ReLU}(1 - x)$$

$$\left. \begin{array}{l} x = 0 \rightarrow f(0) = 0 \\ x = 2 \rightarrow f(2) = 1 \end{array} \right\} \text{Gradient} = 0$$

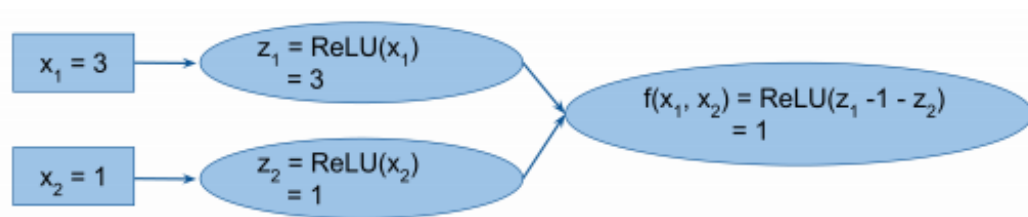


- ✓ Base-line을 사용하는 DeepLIFT (difference from reference)같은 경우 gradient saturate를 막기 때문에 sensitivity 만족

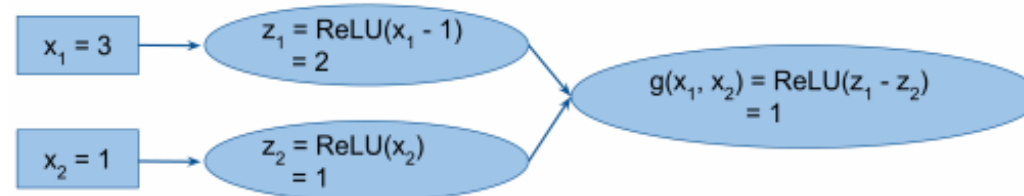
02 | Two Fundamental Axioms

Axioms

- Axiom: Implementation Invariance
 - ✓ 두 모델이 존재할 때, 내부 구현과 상관 없이 동일한 입력에 대해 동일한 출력을 낸다면 (*functionally equivalent*)
입력이 기여하는 정도 또한 두 모델에서 같아야 함
 - ✓ DeepLIFT는 공리를 만족하지 못함



Gradient attributions ($input \times gradient$) : $x_1 = 3, x_2 = -1$
DeepLIFT attributions ($input \times gradient$) : $x_1 = 2, x_2 = -1$



Gradient attributions ($input \times gradient$) : $x_1 = 3, x_2 = -1$
DeepLIFT attributions ($input \times gradient$) : $x_1 = 3, x_2 = -1$

- ✓ 만약 attribution method에서 Implementation Invariance를 만족하지 못하면, 중요하지 않은 feature에 대한 기여도가 민감해질 수 있음 → 중요하지 않은 feature의 기여도가 클 수 있음

03 | Method: Integrated Gradients

- Implementation Invariance와 Sensitivity를 결합한 방법 Integrated Gradients를 제안

✓ $F: \mathbb{R}^n \rightarrow [0,1]$, input $x \in \mathbb{R}^n$, baseline input $x' \in \mathbb{R}^n$

- Baseline x' 부터 input x 까지 straightline path를 그리고, path의 모든 point에 대해 gradient를 구함
- Integrated Gradients는 구해진 gradient를 모두 더한 값 (path integral)
 - Path integral : 이동가능한 경로를 모두 더해서 나타내는 것

$$\text{Integrated Grads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

$\frac{\partial F(x)}{\partial x_i}$ 는 $F(x)$ 의 i^{th} 번 째의 gradient 값

03 | Method: Integrated Gradients

- Axiom: Completeness

- ✓ Integrated gradient는 completeness를 만족함
- ✓ 입력에 대한 기여도는 x 와 x' 의 output의 차이 만큼 더해짐

$$\sum_{i=1}^n \text{Integrated Grads}_i(x) = F(x) - F(x')$$

모든 변수에 대한 grad 합 x 와 x' 의 output의 차이

Additive feature attribution method와 유사

$$F(x) = \sum_i A_i^F(x)$$

- Computing Integrated Gradients

- ✓ Integrated gradient는 Riemman approximation을 통해 추정할 수 있음

Step-size m : 20~300

$$\text{Integrated Grads}_i^{\text{approx}}(x) ::= (x_i - x'_i) \times \sum_{k=1}^m \frac{\partial F \left(x' + \frac{k}{m} \times (x - x') \right)}{\partial x_i} \times \frac{1}{m}$$

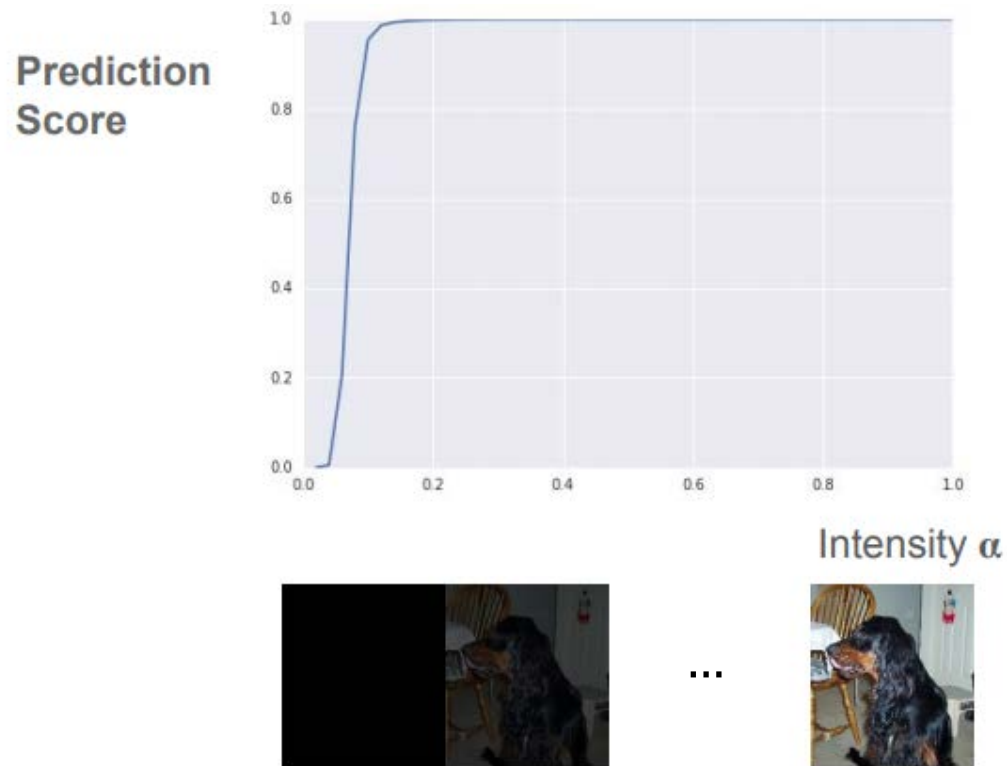
03 | Method: Integrated Gradients

- Attribution using *gradients*

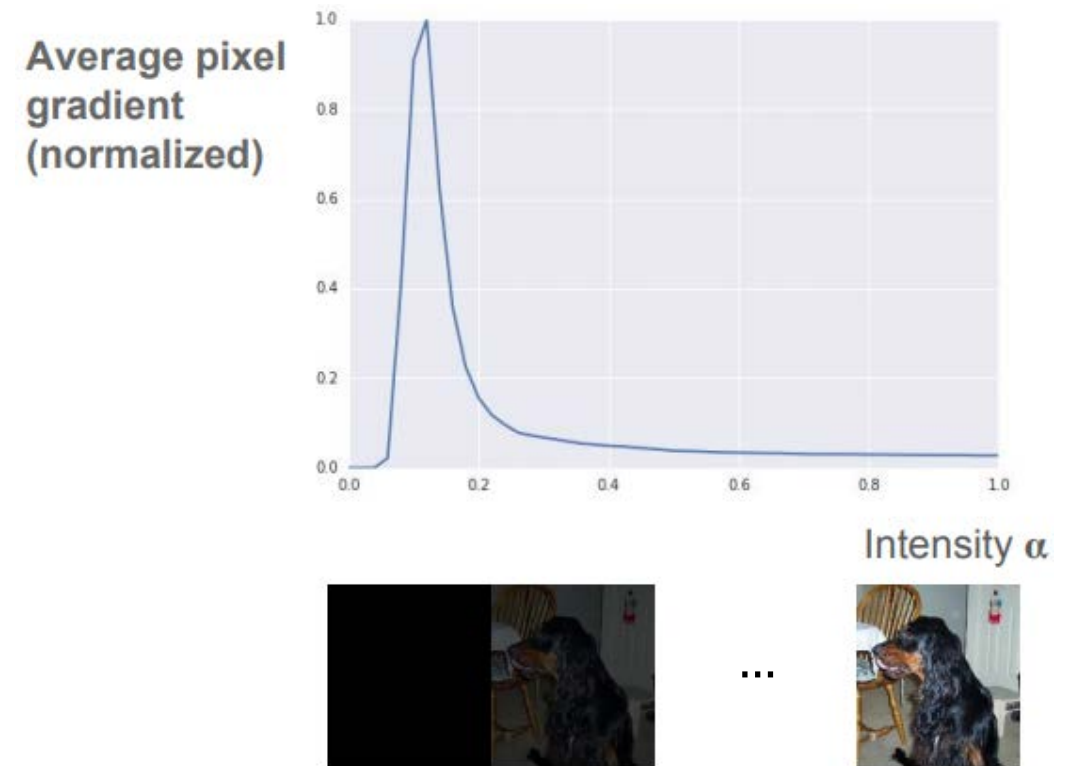


03 | Method: Integrated Gradients

- Attribution using *gradients*



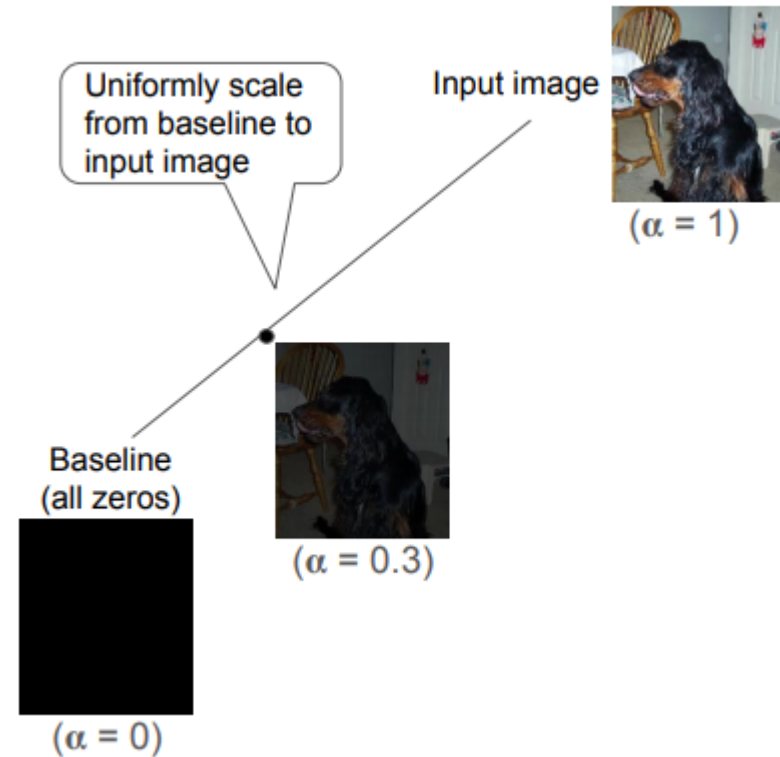
- Attribution using *Integrated gradients*



03 | Method: Integrated Gradients

- Method: Integrated gradients

1. Baseline (black) 이미지부터 실제 이미지까지 이미지의 interpolation을 통해 이미지 셋을 구성
2. 구성된 이미지 셋에 대한 gradients의 평균을 구함



04 | Experiments

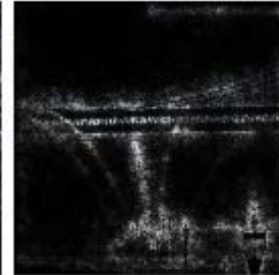


Top label: fireboat

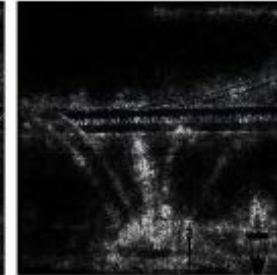
Score: 0.999961



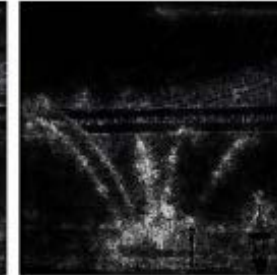
alpha=0.02



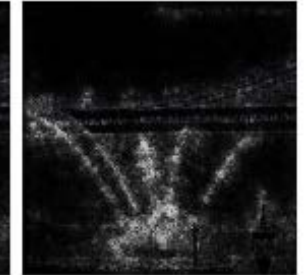
alpha=0.04



alpha=0.06



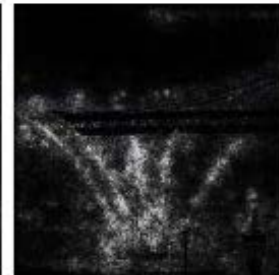
alpha=0.08



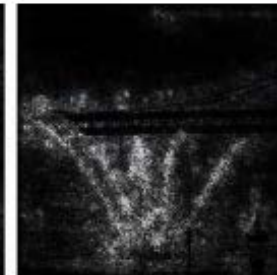
alpha=0.1



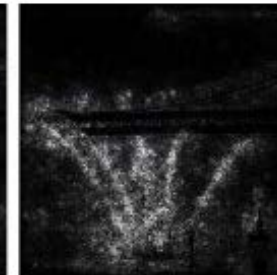
alpha=0.12



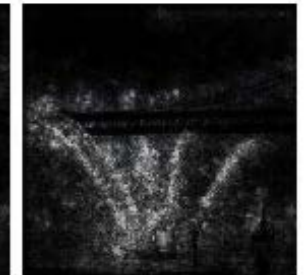
alpha=0.14



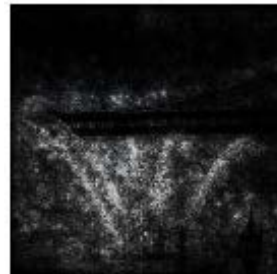
alpha=0.16



alpha=0.18



alpha=0.2



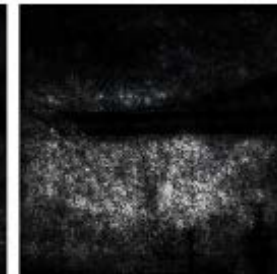
alpha=0.3



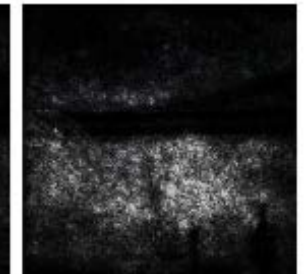
alpha=0.4



alpha=0.6



alpha=0.8



alpha=1.0



04 | Experiments

Original image (Drilling platform)



Gradient at image



Integrated gradient



Original image (Drilling platform)



Gradient at image



Integrated gradient



Q&A

감사합니다.