



Learning Important Features Through Propagating Activation Differences

이상용 / 2020-02-28



Computational Data Science LAB



Learning Important Features Through Propagating Activation Differences

Computational Data Science LAB

목차

1. Introduction
2. The DeepLIFT MEthod
3. Experiments



논의사항 및
결정사항

관련문서

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. 34th International Conference on Machine Learning, ICML 2017, 7, 4844–4866.



CONTENTS

1. Introduction
 2. The DeepLIFT MEthod
 3. Experiments
- 
- 

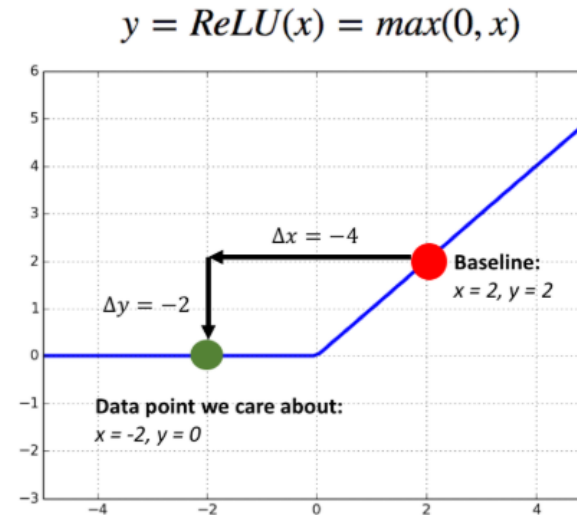
01 | Introduction

- Deep neural networks가 점차 대중화 됨에 따라 black-box 모델의 ‘interpretability’가 중요해짐
- 본 논문은 주어진 *출력에 대해 입력의 중요도 점수를* 할당하는 새로운 알고리즘을 제안
- 제안하는 알고리즘은 DeepLIFT (Deep Learning Important FeaTures)

02 | The DeepLIFT Method

- DeepLIFT의 접근 방식은 개념적으로 매우 간단하지만 구현이 까다로움
 - ✓ DeepLIFT는 y 에 대한 x 의 gradient에 집중하는 것이 아님
 - ✓ ‘reference’ input과 input의 차이와 ‘reference’ output과 output의 차이 (difference from reference)에 초점

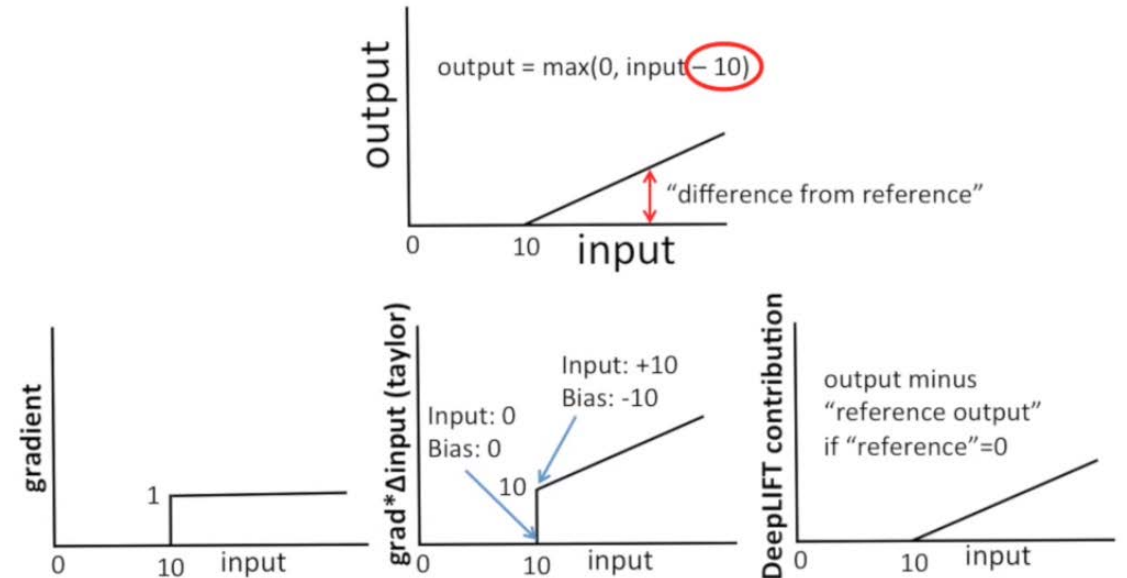
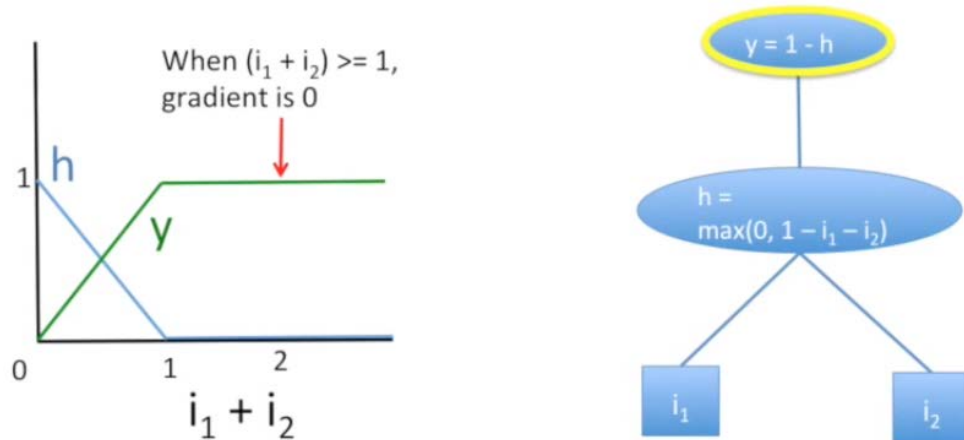
$$slope = \frac{\partial Y}{\partial x_i} \rightarrow \frac{Y - Y^{baseline}}{x_i - x^{baseline}} = \frac{\Delta y}{\Delta x}$$



02 | The DeepLIFT Method

- 기존 방법들의 한계점

$$y = (i_1 + i_2) \text{ when } (i_1 + i_2) < 1 \\ = 1 \text{ when } (i_1 + i_2) \geq 1$$



- ✓ DeepLIFT는 $\frac{\partial Y}{\partial x_i} = 0$ 인 경우에도, non-zero값을 구할 수 있음
- ✓ 또한, bias로 인한 gradient의 sudden jump를 피할 수 있음

02 | The DeepLIFT Method

- Chain rule

- ✓ DeepLIFT는 기존의 backpropagation에서의 chain rule을 적용시킬 수 있음

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial Y} \frac{\partial Y}{\partial x} \rightarrow \frac{\Delta F}{\Delta x} = \frac{\Delta F}{\Delta Y} \frac{\Delta Y}{\Delta x}$$

- ✓ Chain rule을 적용하여 backpropagation을 하며, 모델의 출력에 관련된 입력의 feature importance를 정의 할 수 있음

$$x_i \times \frac{\partial F}{\partial x_i} \rightarrow (x_i - x_i^{baseline}) \times \frac{Y - Y^{baseline}}{x_i - x^{baseline}}$$

기존의 feature importance 방법

Pixel \times gradient \approx Grad-CAM

02 | The DeepLIFT Method

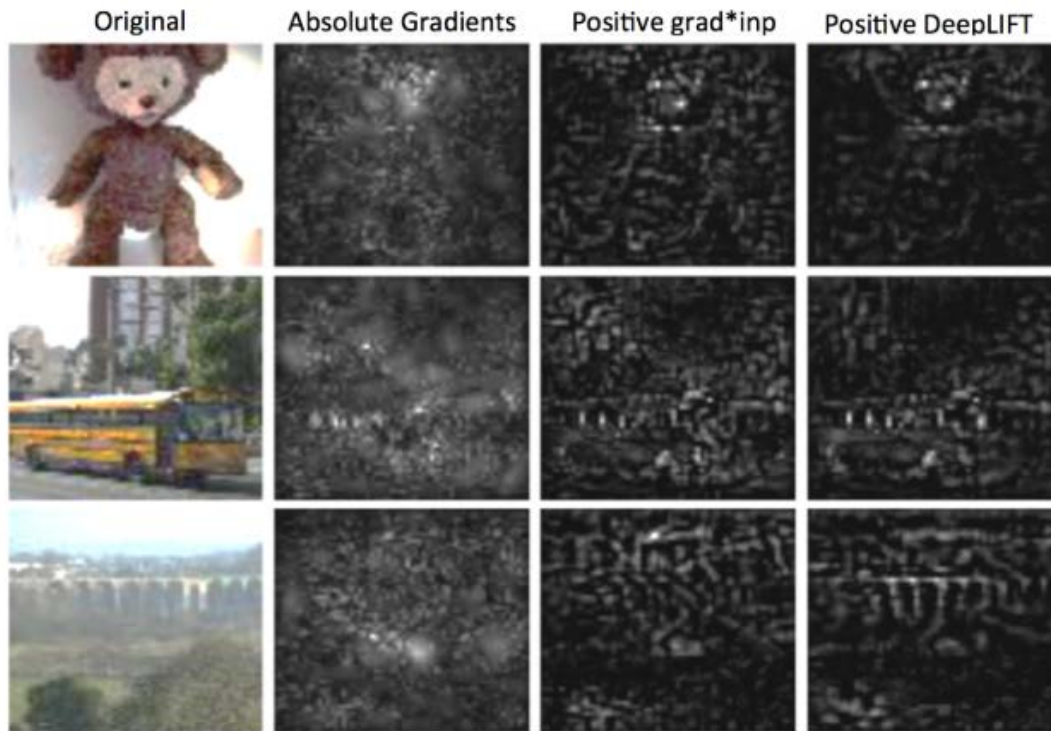
- Defining the Reference

- ✓ DeepLIFT에서 통찰력 있는 결과를 얻기 위해선 reference를 잘 선택해야 함
- ✓ 실제로 좋은 reference를 선택하는 것은 domain-specific knowledge에 의존적임
- ✓ Ex) MNIST: all zeros reference

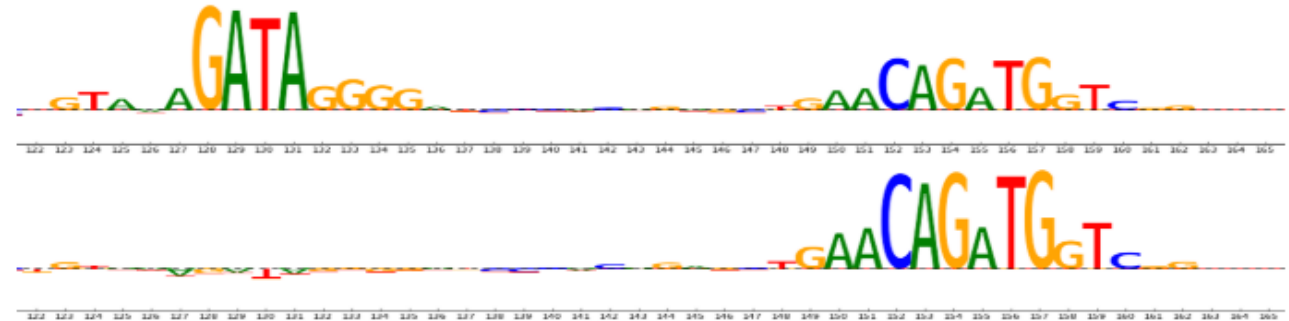
- ✓ 또는, SHAP 라이브러리를 사용하여 편리하게 해석 할 모델의 적절한 기준을 선택할 수 있음

03 | Experiments

- Comparison of methods



- DNA sequence



Q&A

감사합니다.