



On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation



이상용 / 2020-02-24



Computational Data Science LAB



CONTENTS

1. Pixel-wise Decomposition as a General Concept
 2. Layer-wise relevance propagation
 3. Experiments
- 
- 

01 | Pixel-wise Decomposition as a General Concept

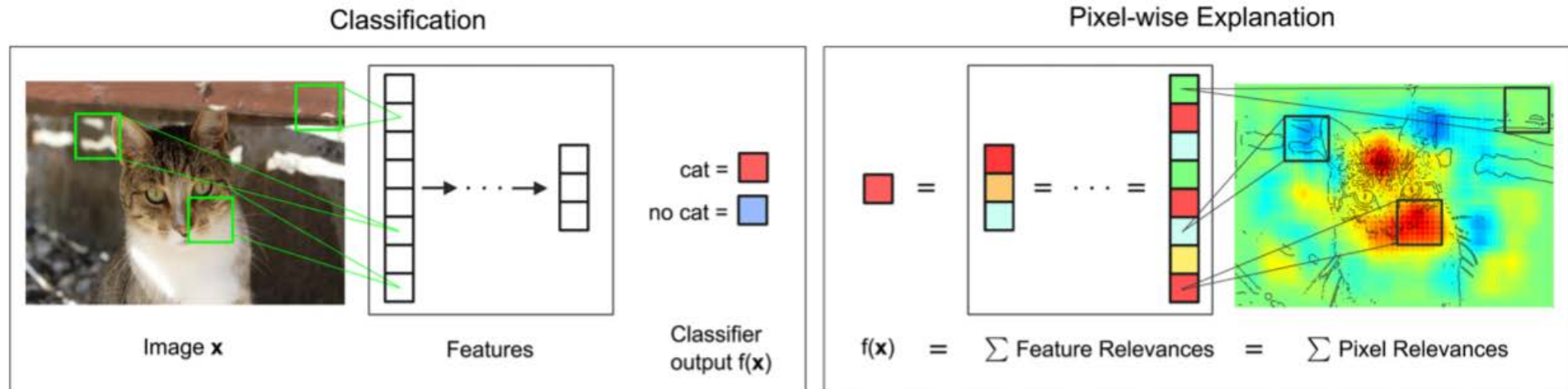
- Introduction
 - ✓ 이 논문은 Classifier가 어떤 결정을 내렸을 때, pixel 단위로 해당 결정에 얼마나 기여를 했는지 분석
- Pixel-wise Decomposition
 - ✓ Pixel-wise Decomposition의 전체적인 아이디어는 classifier f 가 어떤 예측 $f(x)$ 를 했을 때, x 의 변수가 얼마만큼 관련이 있는지 x 를 픽셀 단위로 분해하여 파악하는 것
 - ✓ 어떤 예측 $f(x)$ 를 픽셀단위로 분해하는 방법 중 하나는 각 Input feature의 관련도의 합으로 표현하는 것

01 | Layer-wise relevance propagation

- ✓ 어떤 예측 $f(x)$ 를 픽셀단위로 분해하는 방법 중 하나는 각 Input feature의 관련도의 합으로 표현하는 것

$$f(x) \approx \sum_{d=1}^V R_d \quad \text{모든 Pixel에 대한 관련도의 합은 } f(x) \text{와 유사}$$

- ✓ Main idea



02 | Pixel-wise Decomposition as a General Concept

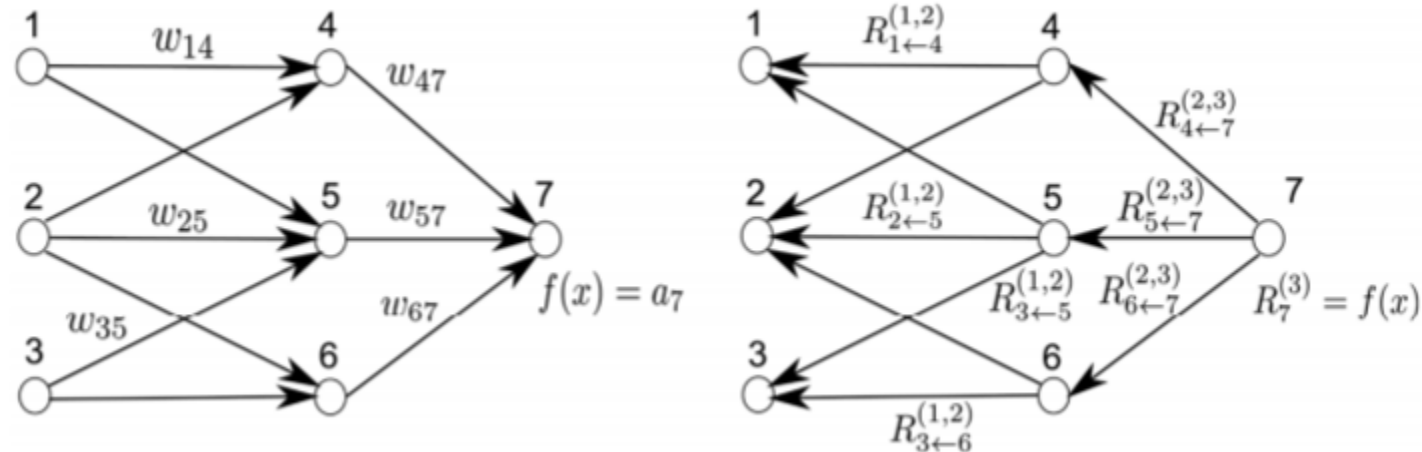
- Layer-wise relevance propagation은 layer별로 관련도가 분해되어 계산할 수 있다고 가정

$$f(x) = \dots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \dots = \sum_d R_d^{(1)}$$

- ✓ 위 식을 반복하면 last layer인 classifier output $f(x)$ 를 픽셀단위의 관련도의 합으로 표현 가능
즉, $f(x) \approx \sum_{d=1}^V R_d$ 만족
- ✓ 위 식은 레이어 사이의 relevance R 의 보존법칙(conservation law)으로 해석할 수 있음

02 | Pixel-wise Decomposition as a General Concept

- Example



$$R_7^{(3)} = R_4^{(2)} + R_5^{(2)} + R_6^{(2)}$$

$$R_4^{(2)} + R_5^{(2)} + R_6^{(2)} = R_1^{(1)} + R_2^{(1)} + R_3^{(1)}$$

02 | Pixel-wise Decomposition as a General Concept

- Example

- ✓ 이 예제는 두 가지 가정을 적용
- ✓ 첫째, 연결된 뉴런 i 와 j 사이의 message $R_{i \leftarrow j}^{(l, l+1)}$ 의 관점에서 layer-wise relevance를 표현 가능
- ✓ 둘째, incoming 메시지의 합으로 output 뉴런 7을 제외한 모든 뉴런의 관련성을 정의 가능

$$R_i^{(l)} = \sum_{k: i \text{ is input for neuron } k} R_{i \leftarrow k}^{(l, l+1)}$$

02 | Pixel-wise Decomposition as a General Concept

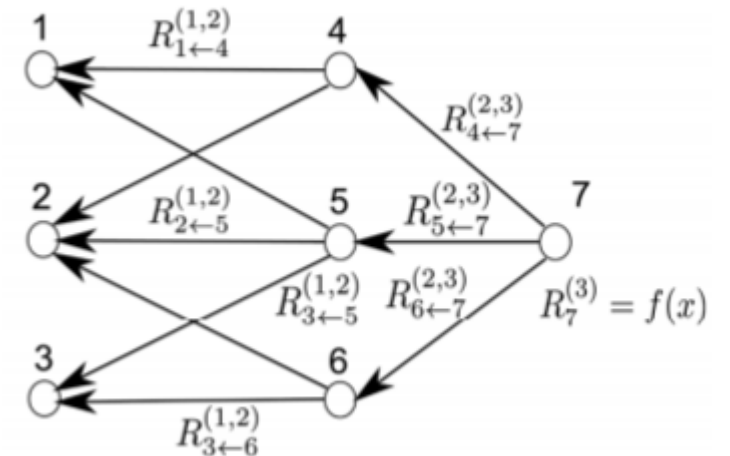
- Example

$$R_7^{(3)} = R_{4 \leftarrow 7}^{(2,3)} + R_{5 \leftarrow 7}^{(2,3)} + R_{6 \leftarrow 7}^{(2,3)}$$

$$R_4^{(2)} = R_{1 \leftarrow 4}^{(1,2)} + R_{2 \leftarrow 4}^{(1,2)}$$

$$R_5^{(2)} = R_{1 \leftarrow 5}^{(1,2)} + R_{2 \leftarrow 5}^{(1,2)} + R_{3 \leftarrow 5}^{(1,2)}$$

$$R_6^{(2)} = R_{2 \leftarrow 6}^{(1,2)} + R_{3 \leftarrow 6}^{(1,2)}$$



✓ 일반식으로 표현 가능

$$\text{특정 뉴런 } R_k^{(l+1)} = \sum_{i: i \text{ is input for neuron } k} R_{i \leftarrow k}^{(l,l+1)}$$

✓ 인풋을 제외한 뉴런 모두 표현 가능

02 | Pixel-wise Decomposition as a General Concept

- Example

$$\begin{aligned} \text{특정 레이어} \quad \boxed{\sum_k R_k^{(l+1)}} &= \sum_k \sum_{i: i \text{ is input for neuron } k} R_{i \leftarrow k}^{(l, l+1)} \\ &= \sum_i \sum_{k: i \text{ is input for neuron } k} R_{i \leftarrow k}^{(l, l+1)} \\ &= \boxed{\sum_k R_i^{(l)}} \end{aligned}$$

✓ 다음 레이어의 공헌도의 합과 현재 레이어의 공헌도의 합이 같음

02 | Pixel-wise Decomposition as a General Concept

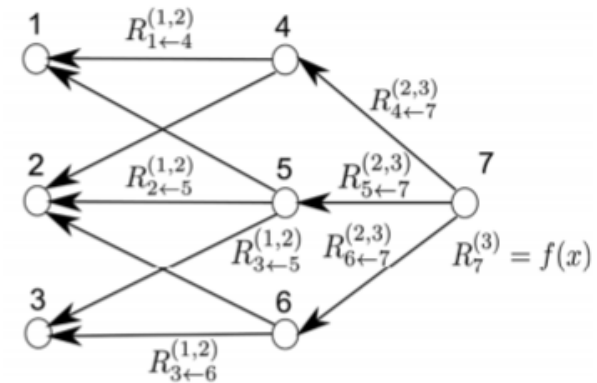
- Example

- ✓ Classification time 동안 뉴런 i 가 뉴런 k 에 인풋인 것을 알고 있으므로, 명백한 form으로 표현 가능

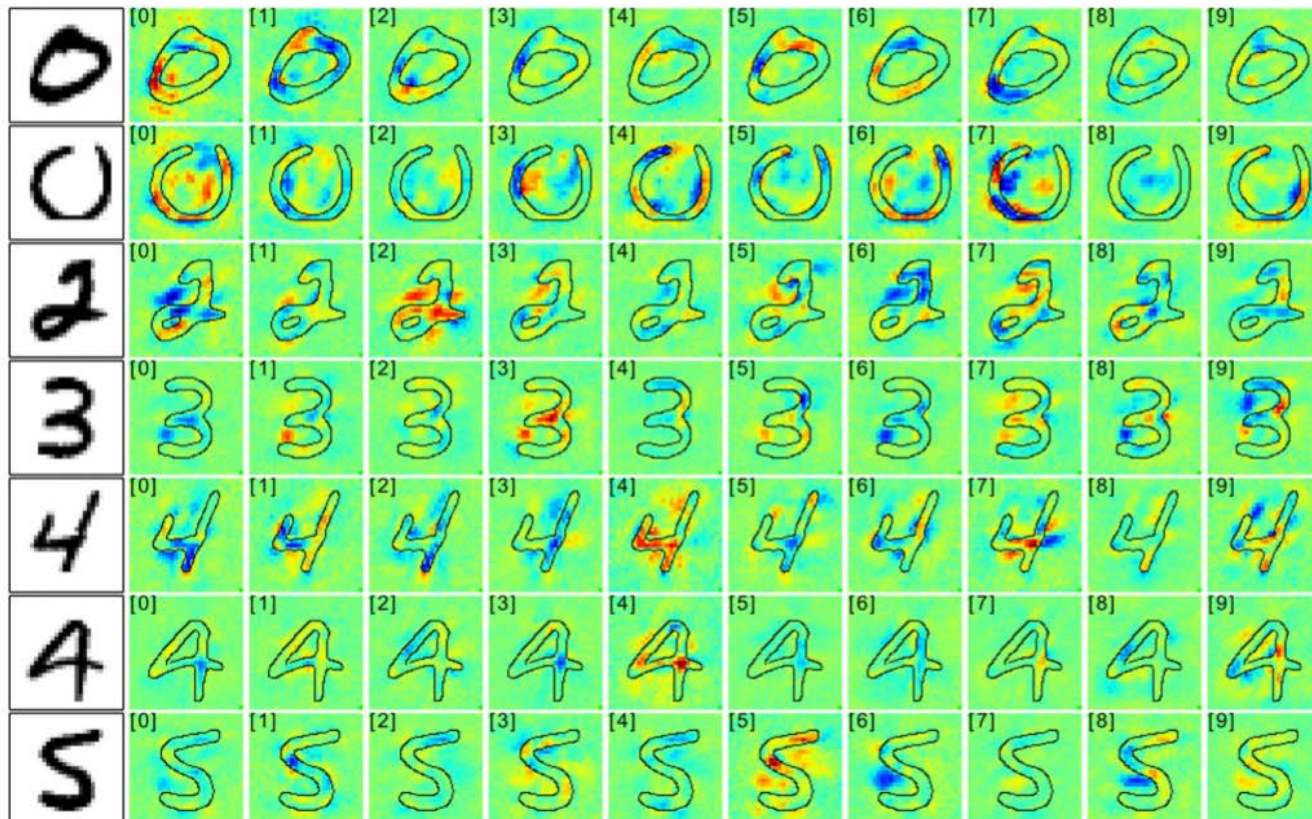
$$R_7^{(3)} = R_7^{(3)} \frac{a_4 w_{47}}{\sum_{i=4,5,6} a_i w_{i7}} + R_7^{(3)} \frac{a_5 w_{57}}{\sum_{i=4,5,6} a_i w_{i7}} + R_7^{(3)} \frac{a_6 w_{67}}{\sum_{i=4,5,6} a_i w_{i7}}$$

- ✓ 일반식으로 표현 가능

$$R_{i \leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}$$



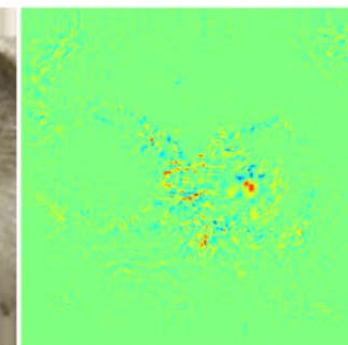
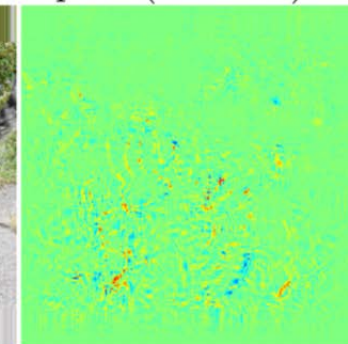
03 | Experiments



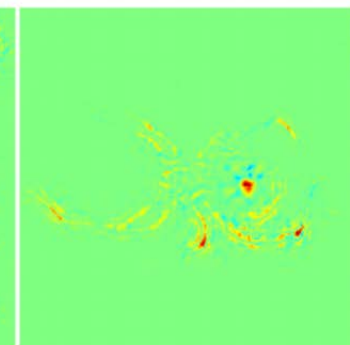
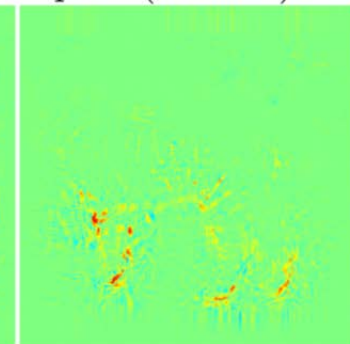
Image



Eq. 58 ($\epsilon = 0.01$)



Eq. 58 ($\epsilon = 100$)



Q&A

감사합니다.