2021/1/15 이유진 (youjin lee)

Paper Review

Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks

Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, Seong-Whan Lee

AAAI, 2020

Abstract

Deep Neural Networks(DNNs)은 다양한 분야에서 우수한 성능을 입증했고, DNNs의 복잡한 내부 메커니즘 이해에 대한 관심이 증가하고 있다. 논문에서 우리는 DNNs의 예측 결과를 층 간 상대적 영향에 따라 관련이 있는지(positive) 없는지(negative) 분류하는 새로운 관점으로 분해하는 Relative Attributing Propagation(RAP)를 소개한다. 각 뉴런의 관련성은 conservation rule을 유지하면서 그것의 기여 정도에 대하여 식별된다. 상대적 우선 순위에 따라 뉴런에 할당된 관련성을 고려하면, RAP는 각 뉴런이 결과와 관련된 양극의 중요도 점수(매우 관련있음부터 매우 무관함까지)로 할당 받도록 한다. 그러므로 RAP는 기존 설명법보다 분리된 속성에 대해 더 명확하고 주의 깊게 시각화 하여 DDNs을 해석할 수 있게 해준다. RAP로 얻은 속성이 각 의미에 대해 정확히 설명되는지 타당성을 입증하기 위해 우리는 평가 지표를 이용했다: (i) Outside-inside relevance ratio, (ii) Segmentation mIOU, (iii) Region perturbation. 모든 실험과 지표에서 RAP는 기존의 방식들에 비해 상당한 차이를 보였다.

Introduction

인상적인 성능에도 불구하고 DNNs의 복잡한 내부 구조로부터 야기된 투명성 문제 때문에 DNNs의 적용은 어려웠다. 따라서 해당 문제를 해결하고자 많은 연구가 최근까지 진행되어 왔고 attributing 방법은 입력층에 관련성 점수를 할당함으로써 입력에 대한 결정에 기여하는 중요한 요인을 드러낸다.

2015년 DNN의 결과에 대한 각 이미지 위치의 positive 및 negative 기여를 고려하기 위해 예측으로부터 관련성을 전파하는 layer-wise relevance propagation(LRP)가 소개되었다. 그러나 기여 정도와 방향을 고려하지 않고 positive 및 negative 기여를 전파하는 것은 해석의 결함이 생길 수 있다. 즉, 복잡한 내부 구조의 구성요소가 이동하고 전달 값이 바뀌기 때문에 출력에 대한 개별 유닛의 실제 영향을 명확히 할 필요가 있다. 게다가 각 뉴런의 관련성은 부호와 관계없이 기여 절대값에 매우 의존적이다.

이 논문에서 우리는 각 뉴런이 연결된 주변과 그 뉴런 간의 영향을 설명하고 그것을 관련 중요도에 할당하여 각 뉴런의 관련성을 해석하는 새로운 관점을 제시한다. 뉴런 간의 영향력에 대한 기여의 sign로부터 관련성의 관점을 바꾸는 것이 주요 아이디어이다. 우리는 conservation을 유지하면서 우선순위를 바꾸고 positive와 negative로 그것을 재조정함으로써 관련성을 재분배한다. 이 방식으로 관련성은 결과의 중요도에 따라 각 뉴런에 방향을 부여하여 할당된다. Fig. 1은 RAP와 기존 방식들을 비교하여 설명한 것이다. 이전 연구가 뉴런의 기여 방향에 근거한 것이라면 우리는 뉴런의 중요도에 따라 관련성을 할당한 것이다. 따라서 우리 연구의 주요 기여는 다음과 같다:

- 우리는 뉴런 간 상대적인 영향력에 따라 각 뉴런에 positive 및 negative 관련성을 부여하는 RAP를 제안한다. 우리는 관련성 의존 현상에 대해 다뤘고 우선순위를 고려한 관련성에 접근하기 위해 새로운 관점의 필요성을 제시했다. 또한 중간층 사이의 실제 기여도에 따라 분류 기준을 정하여 전파하는 동안 degeneracy 위험을 예방했다.
- 우리는 전파된 속성들이 의미 있는지 평가하기 위해 교집합(Outside-Inside relevance ratio, region perturbation)을 적용했다. 이는 RAP로 얻은 속성이 다른 방법들에 비해 무관한 영역을 명확히 분류하면서 높은 객관성 점수를 제공함을 보여주었다.

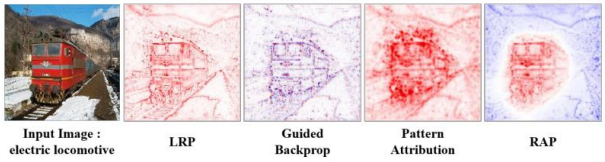

Figure 1: Comparison of the conventional explaining methods and RAP applied to VGG-16. In the previous methods, the attributions are similarly distributed across the entire image. Our RAP clearly distinguishes relevant (red) and irrelevant pixels (blue), placing the relevant attributions on the object, and the irrelevant ones on the background.

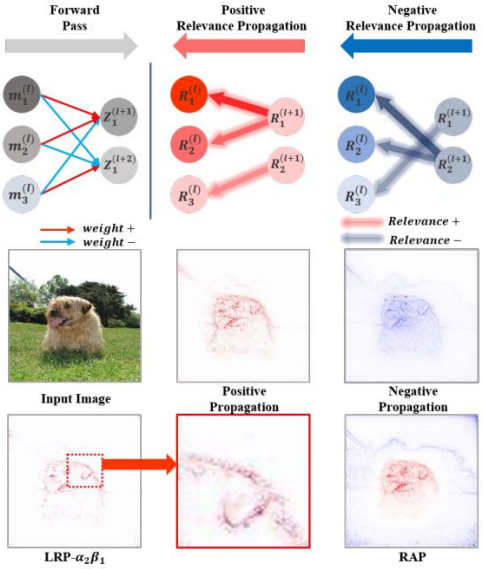

Figure 2: An illustration of the way positive and negative contributions are handled in LRP. See text for details.

Related Work

DNNs이 학습한 것을 이해하기 위한 연구는 최근에도 이어져 오고 있다.

DNN 모델을 해석하려는 관점에서는 은닉층의 활성화를 극대화하여 시각화 하거나 salient feature maps을 생성하는 방식이 있다. (Foerster et al, 2017)은 affine network로 변환된 입력을 소개했고, (Koh and Liang, 2017)은 모델 행동을 이해하기 위해 영향 함수를 제안했다. (Ribeiro & Singh & Guestrin, 2016)은 분류 모델의 예측을 설명하는 알고리즘인 LIME을 제안했다.

DNN의 결정을 설명하려는 관점에서는 픽셀 공간에서 관련성을 재분배하기 위해 입력 기여도를 역으로 전파시켰다. 민감도 분석은 요인을 설명하는 동안 예측 결과에 대한 증거가 되는 DNN에 의해 분류된 입력 이미지의 민감도를 시각화했다(Baehrens et al, 2010). (Zeiler & Fergus, 2014)는 패턴 식별을 위해 deconvolution 방법을 제안했고, (Bach et al, 2015)는 관련성을 역전파하기 위해 LRP를 소개했다. (Samek et al, 2017)는 LRP가 정량적 · 정성적으로 민감도 분석, deconvolution 방법보다 더 나은 설명을 제공함을 입증했다. Guided BackProp (Springenberg et al, 2014)과 Integrated Gradients (Sundararajan & Taly & Yan, 2017)은 각각 단일 및 출력의 평균 편미분 값을 계산한다. Deep Taylor Decomposition (Montavon et al, 2017)은 입력으로부터 기여도 측면에서 뉴런의 활성화를 분해하는 LRP의 연장선이다. DeepLIFT (Shrikumar & Greenside & Kundaje, 2017)는 각 뉴런의 활성화 간 차이를 reference 활성화에 할당하여 분해하는 방식이다. (Ancona et al, 2018)은 앞에 설명된 LRP 이후의 4가지 방식들의 동등성과 근사 조건을 증명했다. (Lundberg & Lee, 2017)은 기존의 설명법을 통합하고 shapley 값을 근사화하여 SHAP 방법을 제안했다.

그러나 negative 관련성을 다루는 데 있어 애매한 시각화의 문제를 분석하는 연구는 없었다. 따라서 우리는 이 문제의 근본적인 원인을 끌어내고 뉴런의 우선순위를 다루는 해결책을 제시하여 관련있는(또는 관련없는) 물체에 대한 명확한 분류를 초래한다.

Background

[Notations]

$f(x)$: 입력 x에 대해 softmax layer를 지나기 전의 출력값
R : 예측 클래스에 해당되는 $f(x)$의 값, attributing 절차에 대한 입력 관련성을 구성
$l+1$번째 층의 뉴런 j는 l번째 층의 뉴런 i로부터 값 $z_{ij}^{(l+1)}$을 받는다.
$m_i^{(l)}$는 번째 층의 활성화 곱으로 얻어진 값이고 가중치는 $w_{ij}^{(l,l+1)}$ 이다.
$m_j^{(l+1)}$는 관련 뉴런에 의해 요약하여 얻은 $z_j^{(l+1)}$에 bias인 $b_j^{(l,l+1)}$를 더하고 활성화 함수 $a(\cdot)$에 적용한 후에 생성된다.
기여도의 positive 및 negative는: $z_{ij} = z_{ij}^+ + z_{ij}^-$는 $z_{ij}^+ = \max(z_{ij}, 0)$이고 $z_{ij}^- = \min(z_{ij}, 0)$이다.

$$z_{ij}^{(l+1)} = m_i^{(l)} w_{ij}^{(l,l+1)}, \qquad z_j^{(l+1)} = \sum_i z_{ij}^{(l,l+1)}, \qquad m_j^{(l+1)} = a(z_j^{(l+1)} + b_j^{(l,l+1)}) \qquad (1)$$

Layerwise Relevance Propagation (LRP)

LRP는 출력에서 입력까지의 관련성 R을 전파시킴으로써 입력에 높은 관련성을 가진 부분을 찾아낸다. 여기서 $R_i^{(l)}$을 l번째 층의 뉴런 i의 관련성이라 하고 $R_j^{(l+1)}$을 $l+1$번째 층의 뉴런 j와 연관된 관련성이라 할 때, 이 conservation은 다음과 같은 형태를 가진다.

$$\sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} \qquad (2)$$

(Bach et al, 2015)는 Eq.2를 만족하는 2개의 관련성 전파 규칙들을 소개했다. 첫번째 규칙은 $LRP-\epsilon$로 불리며 정의하면 다음과 같다. 이에 따르면 l번째 층의 뉴런 i는 $l+1$번째 층의 뉴런의 활성화에 대한 그들의 기여도에 따라 관련성을 받는다. 상수 ϵ는 분모가 0이 되는 경우에 대한 수치적인 불안정을 방지한다.

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_i z_{ij} + \epsilon} R_j^{(l+1)} \qquad (3)$$

두번째 규칙 $LRP-\alpha\beta$는 관련성 전파 과정에서 positive 및 negative 활성화 간 분류가 되는 동안 conservation 원리를 시행한다. $z_{ij} = z_{ij}^+ + z_{ij}^-$임을 기억하고, 전체 관련성을 유지하기 위해서 선택된 파라미터는 $\alpha - \beta = 1$이 성립해야 한다.

$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{z_{ij}^+}{\sum_i z_{ij}^+} - \beta \cdot \frac{z_{ij}^-}{\sum_i z_{ij}^-} \right) R_j^{(l+1)} \qquad (4)$$

The Shortcoming of Current Relevance Propagation Methods

우리는 toy sample을 이용해 현재 LRP에서 관련성이 어떻게 전파되는지 이해한다. Fig. 2에서 2개의 중간층 사이에서의 순방향과 역방향 관련성 전파의 예시를 보여준다. 이를 통해 순방향은 bias를 포함하지 않을 뿐 아니라 배치 정규화도 없음을 알 수 있고 l번째 층의 모든 뉴런들은 non-negative이다. 단순화시키기 위해 모든 가중치의 절대값은 동일하고 어두운 색은 그 뉴런이 높은 값임을 의미한다. Positive 부분의 $R_1^{(l+1)}$은 $\{m_1^{(l)} : m_2^{(l)}\}$의 비율로 뉴런 $m_1^{(l)}$와 $m_2^{(l)}$로 다시 전파된다.

유사하게 negative 관련성 $m_1^{(l)}$도 $R_1^{(l)}$에 대한 관련성이 매우 낮아지도록 초래하는 $R_2^{(l+1)}$의 관련성의 제일 좋은 몫으로 전파 받는다. Eq. 4에서 뉴런 i=1일 때 높은 positive 와 negative 기여를 합하면 이 값은 상쇄된다.

또 다른 관련 현상은 Fig.2의 개 샘플에 설명되어 있다. 관련성은 Eq.4를 사용하여 재귀적으로 전파되어 positive 전파 이미지를 얻는다($\alpha = 1, \beta = 0$). 동일한 과정을 수행하지만 negative 값($\alpha = 0, \beta = 1$와 유사)에 대해서만 negative 전파 이미지를 얻을 수 있다. 이는 같은 위치의 물체가 모두 높은 positive 및 negative 관련성을 받는 것처럼 보인다. 이 때 Eq.4에 의거하여 이들을 합치면 그들은 서로를 상쇄하려는 경향이 있다. 이 positive 및 negative의 결합은 이미지의 3번째 열에 설명되어 있다. 한 쪽이 지배적인 위치를 제외하고는 많은 positive 관련성 값들이 동등하게 큰 negative 값에 의해 상쇄된다. 그러나 확대된 이미지를 살펴보면 이러한 위치는 positive 일수도 negative 일수도 있고 서로 가깝게 나타난다. 그리하여 우리는 각 뉴런의 절대 기여도를 고려하고 관련성을 전파하는 방식을 고려한 새로운 방법인 RAP를 제안한다.

Relative Attributing Propagation

우리의 목표는 상대적으로 층 사이의 영향력에 따라 (안)중요한 뉴런들을 분류하는 것이다. 이 방법은 3단계로 (ⅰ) absolute influence normalization, (ⅱ) propagation of the relevance, (ⅲ) uniform shifting이고, Fig.3에 설명되어 있다.

Absolute Influence Normalization

우선 우리는 마지막 q번째 층에서 예측 노드 j로부터 실제 기여값 $m_i w_{ij}$에 따라 끝에서 두번째인 p번째 층의 뉴런 i에 대한 관련성 값을 전파한다. 왜냐하면 $b_j^{(p,q)}$가 단일 값이기 때문에 각 뉴런의 기여도가 증가함에 따라 이전 p번째 층에 대한 bias $b_j^{(p,q)}$의 관련성을 고려할 수 있다. 아래의 식에 적용시키면 p번째 층의 관련성 값은 positive 및 negative 양쪽 모두로 구성되어 있다.

$$R_i^{(p)} = \left(\sum_i z_{ij}^{+} + \sum_i z_{ij}^{-} \right) * \frac{R_j^{(q)} + b_j^{(p,q)}}{R_j^{(q)}} \tag{5}$$

다음 우리는 절대값 positive와 negative의 비율 $\left|R_i^{(p)+}\right| : \left|R_i^{(p)-}\right|$에 따라 $R_i^{\prime(p)}$을 정규화 한다. $R_i^{\prime(p)}$는 다음 전파에 대한 새로운 입력 관련성이고 출력 층에 대한 상대적인 중요도에 의해 분배된 p번째의 모든 관련성 값이다. 뉴런의 기여도가 커지면 커질수록 더 positive한 관련성이 할당될 것이다. Eq. 6은 첫번째 관련성 전파 과정에서만 적용된다.

$$R_i^{\prime(p)} = \left|R_i^{(p)}\right| * \frac{\sum_i R_i^{(p)}}{\sum_i \left|R_i^{(p)}\right|} \tag{6}$$

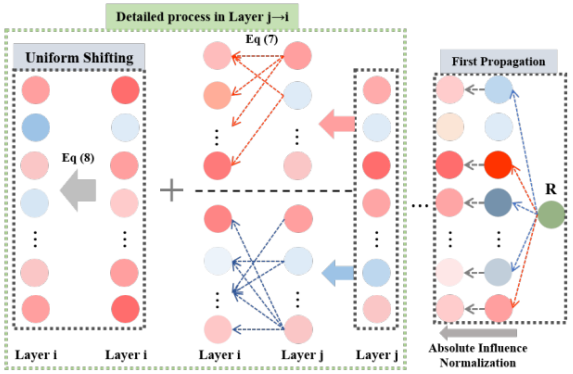

Figure 3: Overall structure of RAP algorithm.

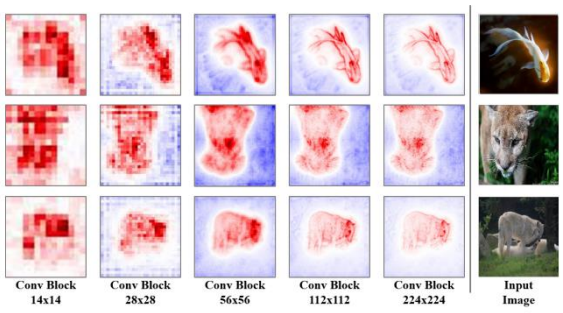

Figure 4: The visualization of the relevance map of the intermediate layers of the VGG-16 network.

Criterion of Relevant Neuron

각 뉴런이 가진 절대적인 기여 정도에 대한 관련도로 바꾼 후, 상대적인 영향도를 유지하면서 관련성을 전파시킬 수 있게 되었다. 우리는 낮은 영향력을 지닌 뉴런들이 negative 관련성을 가지도록 하는 모든 활성화된 뉴런에 uniform shifting을 적용했다. 그 다음 관련성 전파는 positive 기여 $\mathcal{P} = \{i, j | z_{ij}^{+} > 0\}$와 $\mathcal{N} = \{i, j | z_{ij}^{-} < 0\}$를 통해 다음 층에 대한 관련성을 재분배했다. 이는 case $\mathcal{P}$일 경우 positive 영향의 정도에 따라 재분배된 각각의 관련성을 만들어 p번째 층의 관련성을 전파 가능하게 해주고, case $\mathcal{N}$일 경우에도 동일한 절차를 적용한다. 여기서 전파된 관련성은 원래 순방향에서의 negative 전체 기여 간의 비율을 의미한다. 우리는 이 정도의 관련성을 이용하여 활성화된 뉴런 전체를 균일하게 이동시켜 뉴런이 0에 가까운 순서로 negative로 전환되도록 한다.

l번째 층의 각각 관련성 값은 positive 및 negative 기여를 $\{\sum_j z_{ij}^{+} + \sum_j z_{ij}^{-}\}$의 비율로 계산할 수 있다. 그러나 원래 부호로 비율을 계산했을 때 degeneracy 문제가 발생한다. 그러므로 기여 비율은 각 기여의 절대값을 정규화 한 후에 계산된다.

$$v_j^{(l+1)} = R_j^{(l+1)} * \frac{\sum_i |z_{ij}^{-}|}{\sum_i (|z_{ij}^{+}| + |z_{ij}^{-}|)}, \bar{R}_{i \in \mathcal{P},\mathcal{N}}^{(l)} = \sum_j \left(\frac{z_{ij}^{+}}{\sum_i (z_{ij}^{+})} R_j^{(l+1)} + \frac{z_{ij}^{-}}{\sum_i (z_{ij}^{-})} v_j^{(l+1)} \right) \quad (7)$$

negative하게 기여한 뉴런에 대한 관련성을 전파한 후, R_j의 모든 뉴런들은 각 기여의 내부 비율에 따라 관련성 값을 받는다. 그러나 관련성을 보존되지 않으며 이는 $l+1$번째 층에 비해 l번째 층에 과대-할당된다. 원래 negative 기여를 의미하는 과대-할당된 관련성은 모든 활성화된 뉴런들을 균일하게 이동시킴으로써 중요한(중요하지 않은) 뉴런들의 분류에 사용된다. Γ는 i번째 층에서 활성화된 뉴런들의 수를 의미하고, 우리는 이 뉴런 전체에 균등하게 과대-할당된 뉴런들의 평균을 뺀다.

$$\Psi_i^l = \begin{cases} \sum_i \left(\bar{R}_{i \in \mathcal{P},\mathcal{N}}^{(l)} \right) * \frac{1}{\Gamma}, & m_i^{(l)}\text{이 활성화 되어있을 때} \\ 0, & \text{그렇지 않은 경우} \end{cases} \quad (8)$$

이것은 관련성 점수의 일부를 negative 영역으로 이동시킨다. 이 경우 $i \in \mathcal{P}$와 $i \in \mathcal{N}$ 양쪽 그룹 모두에 속하는 뉴런들의 관련성은 아래처럼 된다.

$$R_i^{(l)} = \bar{R}_{i \in \mathcal{P} \cup \mathcal{N}}^{(l)} - \Psi_i^l \quad (9)$$

Eq. 8,9으로 우리는 Eq.2에서처럼 관련성 값이 보존되어 있는지 쉽게 확인했다. 중간층의 feature maps 간에 활성화된 영역이 다르기 때문에 우리는 각 feature map의 중요 영역을 잃지 않기 위해서 균일하게 모든 활성화된 뉴런들에 같은 값을 뺐다. 다음 전파를 위한 negative 입력 관련성은 예측 결과에 상대적으로 낮은 우선순위를 가리킬 것이다. 그러므로 Eq.7은 positive와 negative 방향에서 상대적으로 중요하지 않은 뉴런에 기여한 연결된 뉴런에 negative 관련성을 전파한다.

입력 이미지 층에 대해 마지막 관련성 전파를 위해 우리는 LRP에서 파생된 방식으로 입력층에 전파하는데 일반적으로 사용되는 Z^{β}규칙을 사용했다.

$$R_i^{(0)} = \sum_j \left(\frac{x_i w_{ij} - l_i w_{ij}^{+} - h_i w_{ij}^{-}}{\sum_i (x_i w_{ij} - l_i w_{ij}^{+} - h_i w_{ij}^{-})} R_j^{(l)} \right) \quad (10)$$

우리는 전파 과정동안 relevance map 변화를 조사했고, Fig. 4는 VGG-16 net의 중간 층의 활성화된 뉴런들의 relevance map을 보여준다. (여기서 각 픽셀은 채널 방향 축을 따라 관련성 점수의 합을 나타낸다.) 예상되듯이 positive/negative maps은 점진적으로 분류층(left)에서 입력층(right)으로 바뀐다.

Experimental Evaluations

Qualitative Evaluation of Heatmap

RAP에 의해 생성된 positive 속성을 정성적으로 평가하기 위해 우리는 positive 속성이 수렴하는 영역이 다른 방법들과 얼마나 유사한지 조사함으로써 결과를 비교했다. Fig. 5는 VGG-16 망에 의해 예측된 이미지에 대한 다양한 방법으로부터 생

성된 heatmaps들이고 Fig.6은 ResNet-50의 RAP와 LRP-$\alpha_1\beta_0$을 비교하여 설명했다. negative 속성을 평가하기 위해서 우리는 예측과 관련 없는 부분에 할당된 속성들을 고려했다. 결과적으로 Fig.5을 보면 우리의 결과는 물체와 관련 없는 영역을 명확하게 구별해내지만 다른 방법들은 물체와 해당 영역이 겹쳐 보라색으로 보임을 알 수 있다.

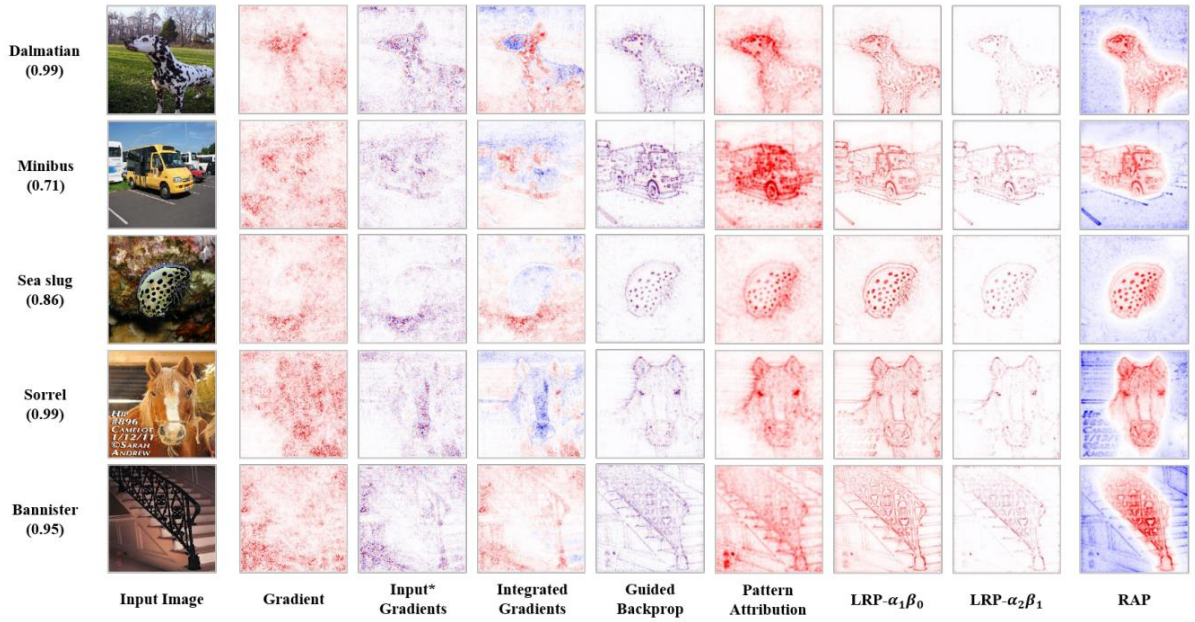

Figure 5: Comparison of the results of conventional methods and our RAP in VGG-16 network.

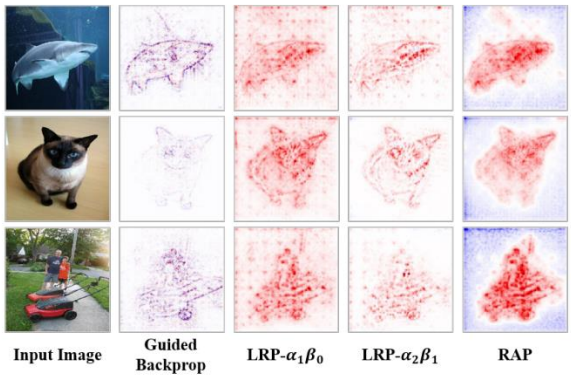

Figure 6: Comparison of visualization results applied on ResNet-50.

Method	mIOU
Guillaumin et al. (Guillaumin, Küttel, and Ferrari 2014b)	57.30
DeepMask (Pinheiro, Collobert, and Dollár 2015)	58.69
DeepSaliency (Li et al. 2016)	62.12
Xiong et al. (Xiong, Jain, and Grauman 2018)	64.22
Ours	62.23

Table 2: Quantitative mIOU results on the ImageNet Segmentation task. Our method is highly comparable to state of the art, despite not using the additional supervision.

Quantitative Assessment of Attributions

DNN 모델을 설명하기 이해 설계된 방법들의 정량적 성능을 평가하는 것은 사소하지 않다. 이 연구에서는 객관성과 관련성을 평가하는데 공통적으로 사용되는 3가지 방법을 사용하여 평가했다: (1) Outside-Inside ratio, (2) Pixel accuracy & Intersection of Union, (3) Region perturbation. (Lapuschkin et al, 2016)은 bounding box의 외부와 내부의 관련성을 계산하여 물체에 속성이 얼마나 초점을 맞추었는지를 평가하는 방법을 소개했다. 우리는 이 방법을 negative 관련성을 제대로 분배했을 때와 아닐 때의 효과를 고려하여 확장했다. (Samek et al, 2017)은 Area over the perturbation curve(AOPC)라 불리는, heatmap으로부터 얻은 픽셀을 계속해서 왜곡하여 영역 변동 과정을 수행하는 방법을 소개했다. 예측과 관련 없는 영역에 분배된 negative 속성을 평가하기 위해 우리는 least relevant first(LeRF)로 작은 변화를 주고 정확도의 저하를 조사했다. 우리는 Imagenet의 validation set으로부터 맞게 분류된 이미지 10,000장을 추출하고 perturbation 테스트 목

적으로 specified bounding boxes를 사용했다. 객관성 점수 측면에서 우리는 segmentation masks를 사용한 Imagenet 데이터셋 이미지 4,276장과 Pascal VOC validation set 이미지 1,449장을 사용했다.

Objectness of Positive Attributions

예측 물체에 대해 분배된 속성들을 검증하기 위해 outside-inside relevance ratio of attributes를 평가했다.

$$\mu = \frac{\frac{1}{|P_{out}|}\sum_{q\in P_{out}} R_q^{(0)+} + \frac{1}{|P_{in}|}\sum_{p\in P_{in}} R_p^{(0)-}}{\frac{1}{|P_{in}|}\sum_{p\in P_{in}} R_p^{(0)+} + \frac{1}{|P_{out}|}\sum_{q\in P_{out}} R_q^{(0)-}} \tag{11}$$

여기서 |·|는 cardinality 연산자이고 $P_{in,out}$는 bounding box의 내·외부 픽셀 집합을 의미한다. positive(negative) 관련성이 bounding box의 외부에 속해 있으면 μ의 값은 증가(감소)된다. 우리는 negative 관련성을 고려하여 결과를 제시했다. Tab.1에서 negative, positive 모두를 고려하면 ALL, negative 관련성을 버릴 때 POS라 한다. 위 표에서 보면 RAP가 두 가지 경우 모두에서 다른 방법들과 비교하여 inside/outside bounding box를 잘 분배하여 최고점을 기록함을 알 수 있고, 나아가 segmentation masks에서도 우수한 성능을 보임을 알 수 있다. 흥미롭게도 Tab.2에 보이듯이 bounding box와 같이 추가적인 supervision으로 학습되는 방법에서 매우 경쟁적이다.

Outside-Inside Ratio		**RAP**	$LRP_{\alpha_1\beta_0}$	$LRP_{\alpha_2\beta_1}$	Gradient	Input* Gradient	Integrated Gradients	Pattern Attribution	Guided Backprop
VGG-16	ALL	**0.252**	-	0.616	-	0.989	1.230	-	1.069
	POS	**0.341**	0.474	0.524	0.619	0.691	0.827	0.415	0.427
Res-50	ALL	**0.164**	-	0.302	-	0.996	1.195	-	1.035
	POS	**0.166**	0.429	0.299	0.597	0.689	0.698	-	0.296
Segmentation Mask		**RAP**	$LRP_{\alpha_1\beta_0}$	$LRP_{\alpha_2\beta_1}$	Gradient	Input* Gradient	Integrated Gradients	Pattern Attribution	Guided Backprop
Imagenet	PIX ACC	**79.23**	75.40	72.95	70.01	66.38	66,52	76.84	71.98
	mIOU	**62.23**	55.78	50.86	49.30	44.01	45.90	58.05	49.87
Pascal VOC	PIX ACC	**73.91**	70.86	69.43	68.14	50.01	52.38	-	66.92
	mIOU	**55.60**	49.82	46.85	46.07	31.69	34.39	-	43.63

Table 1: In Outside-inside relevance ratio result, the first (second) row is the ratio when considering all (only positive) relevance in Imagenet dataset. Segmentation mask result shows the pixel accuracy and mIOU of relevance heatmap in VGG-16 network.

Interfering Negative Attributions

DNN 모델이 올바른 예측을 할 때 예측과 관련 없는 픽셀을 제거한다고 해서 예측 정확도와 관련성 값이 크게 바뀌어서는 안된다. 분류에 영향을 주는 색상과 모양의 왜곡으로 인해 정확도의 감소는 불가피하다. 또한 negative 속성과 일치하는 픽셀을 제거하는 것이 항상 정확도 향상을 가져오는 것은 아님을 인지해야 한다. Fig.7에서 VGG-16과 Resnet-50에 대한 LeRF perturbation을 적용했을 때의 결과를 보여준다. 각 단계마다 negative 관련성에 해당하는 픽셀 100개씩 변화시켜 우리는 총 4,000개의 픽셀을 왜곡했으며, 그 결과 다른 방법과 달리 RAP의 예측 결과는 거의 영향을 받지 않았음을 보인다.

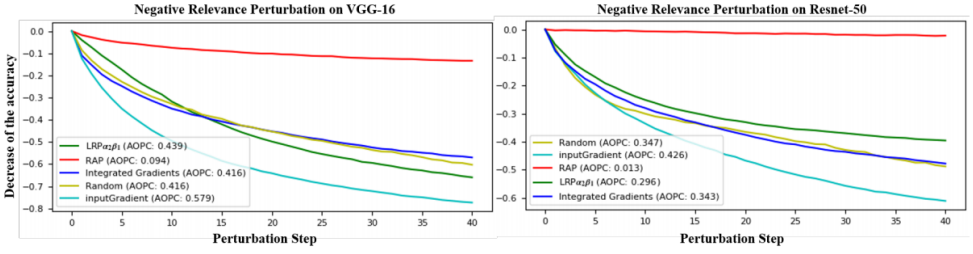

Figure 7: This graph illustrates the results of the negative perturbation on VGG-16 and Resnet-50. For each step, 100 pixels corresponding to the LeRF is perturbated as zero. RAP shows the unique characteristics of the robustness to the perturbation.

Conclusion

이 논문에서 우리는 각 뉴런의 영향력에 따라 관련도(relevance score)를 할당함으로써 DNNs의 예측에 대한 중요성 측면에서 뉴런을 해석하는 RAP를 제안한다. 뉴런 간의 영향력 측면에서 접근하여 예측과 관련 있는 영역과 관련 없는 영역을 구분하는 것이 가능해졌다. 우리는 각 속성들이 유의미한지 평가하기 위해 정량적, 정성적 방식으로 Outside-Inside ratio, mIOU, region perturbation 평가지표를 사용하여 RAP를 평가하였고 결과적으로 다른 방법들에 비해 우수한 성능을 보임을 입증했다. 전반적으로 각 실험은 RAP 방법이 바람직한 특징을 지녔음을 보여주었는데, (1) positive(관련 있음) & negative(관련 없음) 속성들 간의 명확한 구분 (2) 이는 직관적인 객관성을 지니며 또 다른 관련 없는 영역으로부터 이미지 내의 main object를 분류할 수 있다는 것이다.