2020/11/27 이유진 (youjin lee)

Paper Review

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

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015).

PLoS ONE 10 (7): e0130140.

doi:10.1371/journal.pone.0130140

Abstract

자동화된 이미지 분류 시스템의 분류 결정을 이해하고 해석하는 것은 시스템의 추론 과정을 검증하고 전문가에게 추가적인 정보를 제공할 수 있기 때문에 높은 가치가 있다. Machine learning이 많은 과제들을 성공적으로 해결하고 있지만 black box라는 특징 때문에 해당 결정에 도달하게 된 경위에 대한 정보를 제공하지 않는다. 본 논문은 비선형 분류기의 pixel-wise decomposition을 통해 분류 결정을 이해하는 것에 대한 일반적인 해결안을 제안한다. BoW features의 커널 기반 분류기와 다층 neural networks의 예측에 대한 단일 pixel의 기여도를 시각화 할 수 있는 방법론을 소개한다. 이 pixel 기여도는 heatmap으로 시각화 될 수 있으며, 전문가에게 분류 결정의 타당성을 직관적으로 검증할 뿐만 아니라 잠재적인 관심 영역을 제공하여 추가 분석을 가능케한다.

Introduction

이미지 분류 분야에서 Neural Networks, Bag of Words(BoW) models 이 특히 뛰어난 성능을 보이고 있지만, black box이기에 분류기의 예측에 대한 해석 능력이 부족하다는 단점을 지니고 있다. 이는 raw image로부터 모델에 적절한 feature 표현을 위해 다양한 비선형 mapping 과정을 거쳐 최종 classifier function을 통해 결과를 예측하기 때문이다. 위 논문에서는 classification과 interpretability 간의 차이를 좁히는 것을 목표로 (1) BoW는 이미지 내 local feature들의 비선형 mapping의 집합으로써 다루어 질 것이고, (2) neural networks는 일반적인 다층 network 구조로 p-means에 근거한 pooling 기능을 가지는 것으로 생각한다.

Pixel-wise Decomposition as a General Concept

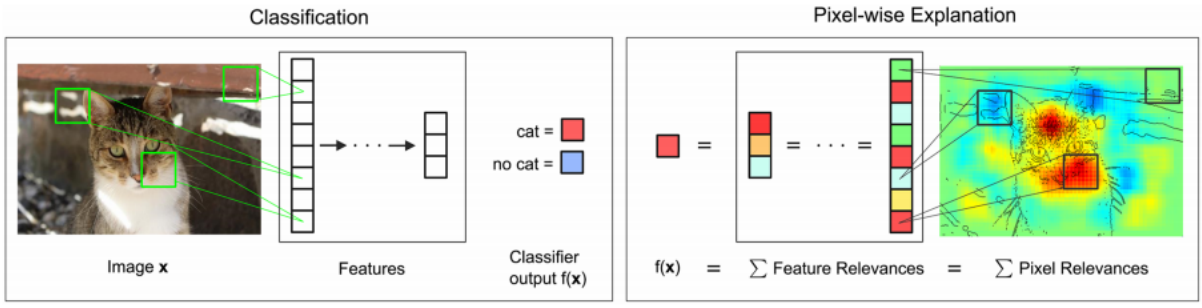

Fig 1. Visualization of the pixel-wise decomposition process. In the classification step the image is converted to a feature vector representation and a classifier is applied to assign the image to a given category, e.g., "cat" or "no cat". Note that the computation of the feature vector usually involves the usage of several intermediate representations. Our method decomposes the classification output $f(x)$ into sums of feature and pixel relevance scores. The final relevances visualize the contributions of single pixels to the prediction. Cat image by pixabay user stinne24.

doi:10.1371/journal.pone.0130140.g001

Pixel-wise Decomposition의 전반적인 아이디어는 이미지 분류 작업에서 분류기 f에 의해 나온 예측 결과 $f(x)$ 에 대해서 이미지 x의 단일 픽셀의 기여도가 어느 정도인지 이해하는 것이다. 분류기가 real-valued outputs이고 threshold가 0에 근사하다고 가정하면, $mapping\ f : \mathbb{R}^V \rightarrow \mathbb{R}^1$가 되고 $f(x) > 0$ 일 때 학습된 구조가 있음을 알 수 있다. 여기서 특정 예측 결과 $f(x)$에 대한 이미지 x 입력 픽셀 x_d의 기여도를 알아내기 위해서 각각의 입력 치수의 합으로써 예측 결과 $f(x)$를 분해하는 방법이 있다.

$$f(x) \approx \sum_{d=1}^{V} R_d \qquad (1)$$

위 식에 근거하여 정성적 해석 측면에서 $R_d < 0$이면 구조의 부재를, $R_d > 0$이면 구조가 있음을 의미한다. 본 논문에서는 식(1)과 같이 pixel-wise decomposition을 하기 위해 일반적인 개념으로서 layer-wise relevance propagation이라고 정의한다. 또한 layer-wise relevance propagation의 근사치를 산출하는 taylor 분해에 근거한 접근법에 대해 논의한다. 위 방법은 학습시에 사전 pixel-wise labeling을 필요로 하지 않지만, pre-trained classifier 위에 구축된다.

■ Layer-wise relevance propagation(LRP)

제약 조건에 의해 정의된 개념으로, 해당 조건을 만족하는 모든 해결안은 LRP의 개념을 따르는 것으로 간주될 것이다. 일반적인 형태의 LRP는 분류기가 여러 개의 연산 레이어로 분해될 수 있다고 가정하고, 이 레이어는 이미지로부터 추출된 feature 부분이거나 feature에 의해 실행된 분류 알고리즘 부분일 수 있다.

첫번째 레이어는 입력(이미지의 픽셀), 마지막 레이어는 분류기 f의 실제 예측 값이다. l-번째 레이어는 벡터 $z = (z_d^{(l)})_{d=1}^{V(l)}$ 로 표현하고, LRP는 우리가 $l+1$번째 레이어에서 벡터 $z_d^{(l+1)}$ 의 relevance score $R_d^{(l+1)}$을 가진다고 가정한다. 이 아이디어는 입력 계층에 더 가까운 l 번째 레이어에서 벡터 $z_d^{(l)}$ 의 각 차원에 대한 relevance score $R_d^{(l)}$ 를 찾는 것이다.

$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)} \qquad (2)$$

식(2)를 반복하여 output 레이어부터 input 레이어까지 따라가면서 식(1)을 산출한다.

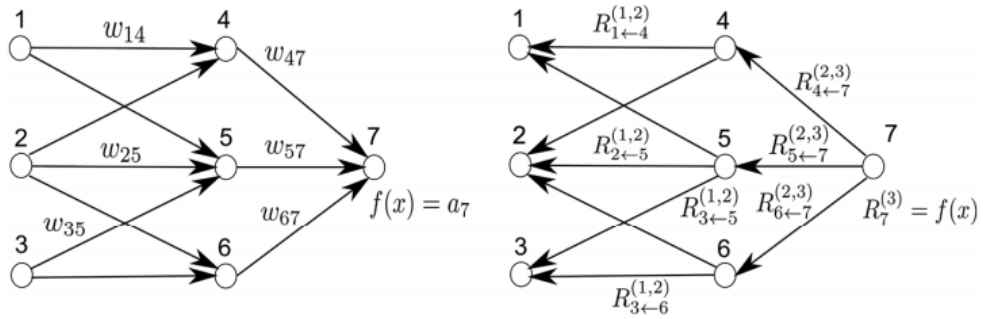

Fig 2. Left: A neural network-shaped classifier during prediction time. w_{ij} are connection weights. a_i is the activation of neuron *i*. Right: The neural network-shaped classifier during layer-wise relevance computation time. $R_i^{(l)}$ is the relevance of neuron *i* which is to be computed. In order to facilitate the computation of $R_i^{(l)}$ we introduce messages $R_{i\leftarrow j}^{(l,l+1)}$. $R_{i\leftarrow j}^{(l,l+1)}$ are messages which need to be computed such that the layer-wise relevance in Eq (2) is conserved. The messages are sent from a neuron *i* to its input neurons *j* via the connections used for classification, e.g. 2 is an input neuron for neurons 4, 5, 6. Neuron 3 is an input neuron for 5, 6. Neurons 4, 5, 6 are the input for neuron 7.

doi:10.1371/journal.pone.0130140.g002

설명에 앞서 2가지 가정을 한다. 첫번째, 뉴런 i와 j사이의 각 연결에 따라 전송될 수 있는 메시지 $R_{i\leftarrow j}^{(l,l+1)}$ 의 layer-wise relevance를 표현한다. 그러나 메시지는 fig2의 오른쪽 그림에 보이듯이 예측 과정과는 대조적인 방향으로 향한다. 두번째, 7번 뉴런을 제외한 모든 뉴런의 relevance를 들어오는 메시지의 합으로 정의한다.

$$R_i^{(l)} = \sum_{k:i\ is\ input\ for\ neuron\ k} R_{i\leftarrow k}^{(l,l+1)} \qquad (3)$$

예를 들어 $R_3^{(1)} = R_{3\leftarrow 5}^{(1,2)} + R_{3\leftarrow 6}^{(1,2)}$ 이다. 단, 7번 뉴런의 relevance는 $R_7^{(3)} = f(x)$ 이다. 여기서 1,2번 뉴런은 inputs이고 뉴런4번의 source인 반면, 6번 뉴런은 2,3번 뉴런의 sink이다.

$$R_i^{(l+1)} = \sum_{k:i\ is\ input\ for\ neuron\ k} R_{i\leftarrow k}^{(l,l+1)} \qquad (4)$$

(3)과 (4)의 차이는 (4)의 합은 $l+1$ 레이어에서 고정 뉴런 k에 대해 l 레이어의 source로 통과하고, (3)의 합은 합은 l 레이어에서 고정 뉴런 k에 대해 $l+1$ 레이어의 source로 통과한다는 것이다. 다음 절에서 (3)과 (4)를 LRP를 정의하는 주요 제약조건으로 설정한다. 이 개념을 따른 해결안은 이 방정식들에 따라 $R_{i\leftarrow k}^{(l,l+1)}$ 를 정의하는데 필요하다.

LRP는 분류하는 동안 전달된 메시지를 반영해야 한다. 이 때 우리는 뉴런 i가 뉴런 k에 입력됨을 알고 있으므로 아래와 같이 방정식을 다시 쓸 수 있다.

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

즉, 위 식을 다시 일반적으로 표현하면 아래의 방정식과 같다.

$$R_{i\leftarrow k}^{(l,l+1)} = R_k^{(l+1)}\frac{a_i w_{ik}}{\sum_h a_h w_{hk}} \qquad (6)$$

(6)은 분모가 0이 될 때 적용되어야 하지만, 어떤 메시지 $R_{i\leftarrow k}^{(l,l+1)}$가 주어질 수 있는지 선행 레이어 l로 부터 뉴런 i의 입력에 비례하여 가중치로 제공한다. 이는 다른 분류 아키텍처를 사용하고 주어진 레이어에서 feature 벡터의 차원으로 뉴런의 개념을 대체할 때 유사한 방식으로 유지된다.

요약하자면, 우리는 feed-forward network 내에서의 layer-wise relevance propagation에 대해 소개했다. 우리가 제안한 정의에서 총 relevance는 1개의 레이어에서 다른 레이어로 유지되도록 제한되며, 총 node relevance는 해당 node로 들어오는 모든 relevance의 합과 같아야 한다. 해당 정의는 해결안으로 주어지지는 않지만 해결안이 충족해야 할 제약 조건으로 주어진다.

■ Taylor-type decomposition

(1)과 같이 일반적인 미분가능 predictor f에 대한 분해를 하기 위한 대안은 1차 taylor 근사법이다.

$$\begin{aligned} f(x) &\approx f(x_0) + Df(x_0)[x - x_0] \\ &= f(x_0) + \sum_{d=1}^{V} \frac{\partial f}{\partial x_{(d)}}(x_0)\left(x_{(d)} - x_{0(d)}\right) \end{aligned} \qquad (7)$$

위에서 말했듯이 분류의 경우 $f(x)$ 값의 부호는 학습된 구조의 존재 유무를 나타내기에 우리는 $f(x_0) = 0$인 집합의 예측 최대 불확실성에 대한 각 픽셀의 기여도를 알아내는 것에 관심있다. 따라서 x_0는 predictor f의 루트가 되도록 선택되어야 한다. 또한, 예측의 taylor 근사값에 대한 정확성을 위하여, x_0는 Euclidean norm 내에서 x에 근접하도록 선택되어야 한다. (7)을 간단히 정리하면 아래와 같이 나온다. $x_{(d)}$가 변했을 때 $f(x)$는 얼마

나 변했는가를 계산하는 식이다.

$$f(x) \approx \sum_{d=1}^{V} \frac{\partial f}{\partial x_{(d)}}(x_0)\left(x_{(d)} - x_{0(d)}\right) \quad such\ that\ f(x_0) = 0 \qquad (8)$$

Fig3은 local gradients(검은 화살표)와 the dimension-wise decomposition(빨간 화살표)의 질적 차이점에 대해 설명한다.

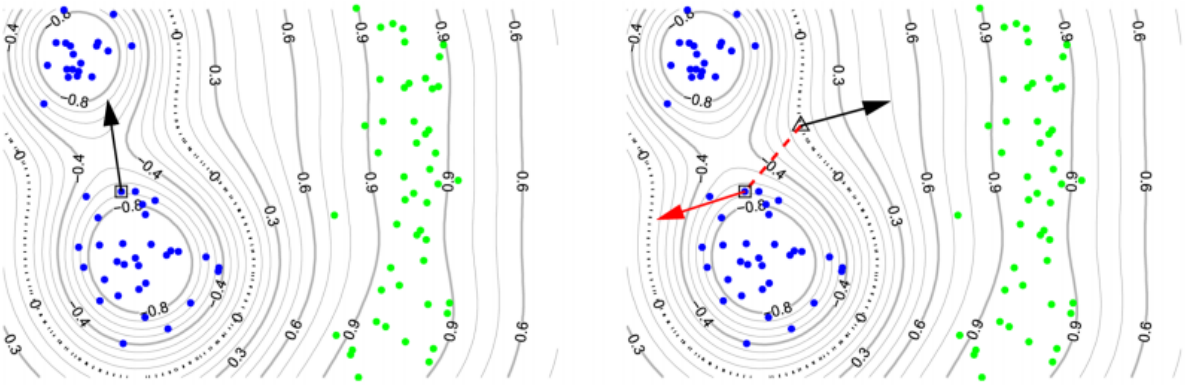

Fig 3. An exemplary real-valued prediction function for classification with the dashed black line being the decision boundary which separates the blue from the green dots. The blue dots are labeled negatively, the green dots are labeled positively. Left: Local gradient of the classification function at the prediction point. Right: Taylor approximation relative to a root point on the decision boundary. This figure depicts the intuition that a gradient at a prediction point *x*—here indicated by a square—does not necessarily point to a close point on the decision boundary. Instead it may point to a local optimum or to a far away point on the decision boundary. In this example the explanation vector from the local gradient at the prediction point *x* has a too large contribution in an irrelevant direction. The closest neighbors of the other class can be found at a very different angle. Thus, the local gradient at the prediction point *x* may not be a good explanation for the contributions of single dimensions to the function value $f(x)$. Local gradients at the prediction point in the left image and the Taylor root point in the right image are indicated by black arrows. The nearest root point x_0 is shown as a triangle on the decision boundary. The red arrow in the right image visualizes the approximation of $f(x)$ by Taylor expansion around the nearest root point x_0. The approximation is given as a vector representing the dimension-wise product between $Df(x_0)$ (the black arrow in the right panel) and $x - x_0$ (the dashed red line in the right panel) which is equivalent to the diagonal of the outer product between $Df(x_0)$ and $x - x_0$.

doi:10.1371/journal.pone.0130140.g003

* relevance : 입력의 변화에 따른 출력의 변화 정도

Taylor-type decomposition은 1개의 레이어 또는 레이어들의 부분집합에 적용될 때, 함수가 매우 비선형일 때 대략적인 relevance propagation 방법으로 볼 수 있다. 특히, 출력 레이어의 relevance가 예측 함수 $f(x)$의 값으로 초기화되어 (6)이 대략적으로 (2)를 만족할 때 적용된다. Taylor 근사와 달리 LRP는 입력값 외의 두 번째 point를 사용할 필요가 없기에 BoW 및 neural network에서 taylor 확장을 통해 근사할 필요없이 광범위한 아키텍처에서 LRP가 구현될 수 있다.

■ Related Work

여러 연구들은 특히 neural network, 커널 기반 분류기, BoW 분류기를 설명하는 주제에 전념해왔다. Neural network의 경우, 픽셀 단위에도 적용되는 뉴런의 결정을 분석하는데 집중했고 이는 CNN에서 특정되어 연구되어왔다(Hansen K, Baehrens D, Schroeter T, Rupp M, Müller KR. Visual Interpretation of Kernel-Based Prediction Models). 이전의 연구들과는 달리 우리는 레이어를 거치면서 relevance 값을 보존하는 것을 목적으로 상위 레이어의 relevance에 가중치를 부여하기위해 하위 레이어로부터 나온 뉴런의 signed activation를 사용한다. 이는 neural network의 구조에 대한 추정을 인코딩하는데 사용된다.

다른 접근 방법은 입력값 x에 대한 편도함수와 x_0을 중심으로 한 full taylor 급수 사이에 존재한다.(Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps). 해당 연구는 도함수를 계산하기 위해 입력값 x과 다른 값 x_0와 taylor 급수의 full linear weighting을 사용하기 위해 remainder bias를 사용한다. 이는 gradient ascent에 의해 해결가능한 최적화 문제를 통해 뉴런의 활성

화를 최대화하는 입력값을 찾으려고 노력한다. 반면 우리의 접근 방식은 특정 뉴런의 최적값을 찾는 것이 아니라 주어진 입력에 대한 결정을 설명하는 것을 목표로 한다.

결과적으로 선행 연구들과 달리 본 연구는 neural network와 Bag of Words features에 대한 layer-wise relevance propagation에 대해 소개한다.

Pixel-wise Decomposition for Classifiers over Bag of Words Features

Neural networks의 진보에도 불구하고, BoW는 여전히 이미지 분류 분야에서는 인기있는 모델이다. 과거 Pascal VOC, ImageCLEF Photo Annotation과 같은 대회에서 우수한 성능을 보였기 때문이다. 경험상 BoW 모델은 샘플 크기가 작은 작업에서는 잘 수행되는 반면, neural network는 richer parameter structure로 인해 overfitting될 위험이 있다.

■ Bag of Words models revisited

BoW feature들을 local features들의 비선형 mapping의 집합으로 간주할 것이고, 모든 BoW 모델들은 hierarchical clustering, fisher vectors 등 기반과 관계없이 공통적으로 다단계 절차를 공유한다.

1번째 단계에서 SIFT와 같은 local feature는 이미지의 작은 영역에 걸쳐 계산되며, 이는 이미지의 모양 특징, 색상, 질감 등과 같은 정보를 함축한다. 학습 중 한 번 수행되는 2번째 단계에서 local feature 공간의 representatives이 계산되고, 이 representatives 집합은 이미지가 벡터로 설명될 수 있도록 돕는 어휘 역할을 한다. 3번째 단계에서는 representatives 집합에 대한 상대적인 local feature들의 통계를 계산한다. 이 통계 수치들은 일반적으로 sum 또는 max-pooling에 의해 수행된 BoW representation x를 산출하기 위해, 이미지 내의 모든 local feature l 에서 집계된다.

$$x_{(d)} = (M^{-1} \sum_{j=1}^{M} \left(\left(m_{(d)}(l_j) \right)^p \right)^{\frac{1}{p}} \qquad (9)$$

m은 mapping function이고, $m_{(d)}$는 d차원에 대한 mapping을 의미한다. 특별한 경우의 sum 과 max-pooling에서 p = 1이고 $limit\ p = \infty$를 포함한다. 마지막으로 커널 기반의 분류기는 이러한 feature들 위에 적용된다. 간단히 하기 위해 우리는 SVM을 활용하면 아래의 식과 같이 나온다.

$$f(x) = b + \sum_{i=1}^{s} \alpha_i y_i k(x_i, x) \qquad (10)$$

이러한 가정은 다중 커널 함수를 사용하는 접근법에 대한 일반성을 잃지 않고 확장될 수 있다.

$$f(x) = b + \sum_{i=1}^{S} \sum_{u=1}^{K} \alpha_{i,u} k_u \left(x_{i(u)}, x_{(u)} \right) \qquad (11)$$

■ Overview of the decomposition steps

이 절에서는 개별 local features 및 최종 단일 픽셀들의 커널 기반 분류 예측 기여도의 decomposition를 도출할 것이다. 해당 접근법은 총 3단계로 구성된다.

(1) kernel 유형에 따라 taylor-type decomposition 이나 layer-wise relevance propagation을 사용한다. 즉, 모든 차원 x의 개별적인 예측의 합으로 contribution $R_d^{(3)}$ 얻는다.

- sum decomposable kernels : layer-wise relevance propagation

$$R_d^{(3)} = \frac{b}{V} + \sum_{i=1}^{S} \alpha_i y_i k^{(d)}(x_{i(d)}, x_{(d)}) \quad (12)$$

- differentiable kernels : taylor-type decomposition

$$R_d^{(3)} \doteq (x - x_0)_{(d)} \sum_{i=1}^{S} \alpha_i y_i \frac{\partial k(x_i, \cdot)}{\partial x_{(d)}}(x_0) \quad (13)$$

(2) relevance score $R_d^{(3)}$로부터 local feature l에 대한 relevance score $R_d^{(2)}$를 얻기 위해 layer-wise propagation을 적용할 것이다.

- local feature scores for sum pooling

$$R_l^{(2)} \doteq \sum_{d \notin Z(x)} R_d^{(3)} \frac{m_{(d)}(l)}{\sum_{l'} m_{(d)}(l')} + \sum_{d \in Z(x)} R_d^{(3)} \frac{1}{|\{l'\}|} \quad (14)$$

- local feature scores for p-means pooling

$$R_l^{(2)} \doteq \sum_{d \notin Z^{(p)}(x)} R_d^{(3)} \frac{m^p{}_{(d)}(l)}{\sum_{l'} m^p{}_{(d)}(l')} + \sum_{d \in Z^{(p)}(x)} R_d^{(3)} \frac{1}{|\{l'\}|} \quad (15)$$

(3) 동일하게 relevance score $R_d^{(2)}$로부터 local feature l에 대한 relevance score $R_d^{(1)}$를 얻고, color-coding에 의해 heatmaps으로 시각화한다.

$$R_q^{(1)} = \sum_{l \in L(q)} \frac{R_l^{(2)}}{|area(l)|} \quad (16)$$

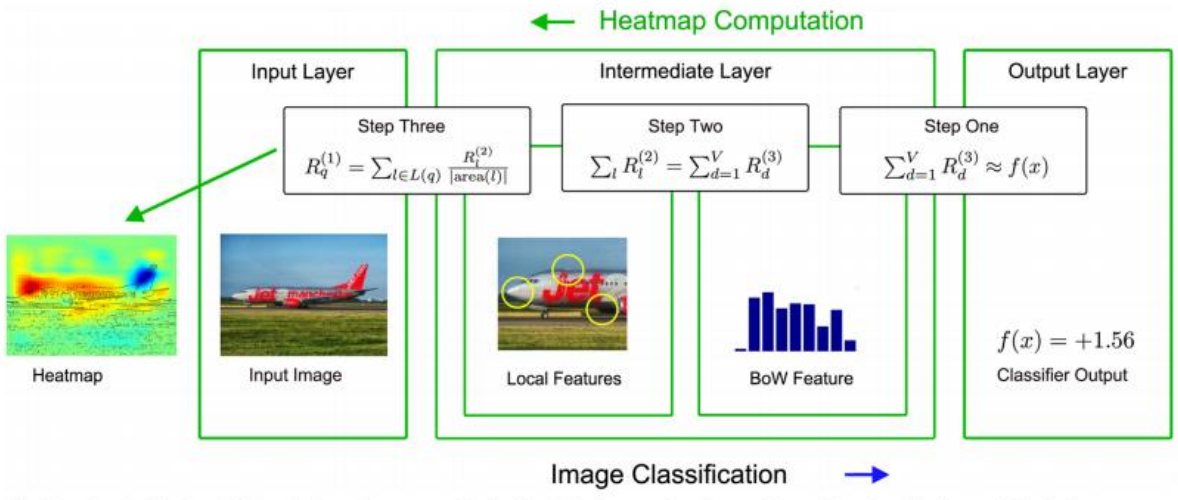

Fig 4. Local and global predictions for input images are obtained by following a series of steps through the classification- and pixel-wise decomposition pipelines. Each step taken towards the final pixel-wise decomposition has a complementing analogue within the Bag of Words classification pipeline. The calculations used during the pixel-wise decomposition process make use of information extracted by those corresponding analogues. Airplane image in the graphic by Pixabay user tpsdave.

doi:10.1371/journal.pone.0130140.g004

Pixel-wise Decomposition for Multilayer Networks

Multilayer networks는 일반적으로 레이어로 구성된 상호 연결된 뉴런 집합으로 구축된다.

$$z_j = \sum_i Z_{ij} + b_j \qquad (17)$$

Neural network의 공식화는 단순한 다층 퍼셉트론, CNN과 같은 광범위한 아키텍처를 포괄할 만큼 일반적이다.

■ Taylor-type decomposition

Network의 입력값과 결과값 간 mapping을 수행하는 벡터값 다변량 함수라고 했을 때, 분류 결정에 대해 우선적으로 가능한 설명은 결정 함수 f 의 root point x_0 에서 taylor 확장에 의해 얻을 수 있다 (13). Backpropagation 알고리즘을 이용한 network topology를 재사용하여 pixel-wise decomposition에 필요한 도함수 $\partial f(x)/\partial x_{(d)}$를 계산할 수 있다.

$$\frac{\partial f}{\partial x_i} = \sum_j \frac{\partial f}{\partial x_j} \cdot \frac{\partial x_j}{\partial x_i} = \sum_j \frac{\partial f}{\partial x_j} \cdot w_{ij} \cdot g'(z_j) \qquad (18)$$

Taylor 기반 decomposition의 조건은 x 에 대한 분류 결정에 대한 local 설명을 지원하는 root point x_0를 찾는 것이다. 이 root point는 x의 이웃에서의 local search을 통해서 또는 x 로 정의된 segment와 다른 클래스의 가장 가까운 이웃에 대한 line search를 통해 찾을 수 있다. 그러나 line search를 통해 찾을 경우, 데이터가 sparse하면 pixel-wise decomposition과 동떨어진 root point x_0를 생성한다.

■ Layer-wise relevance backpropagation

Taylor-wise decomposition의 대안으로 역방향에서 각 레이어의 relevance를 계산하는 것이 가능하다.

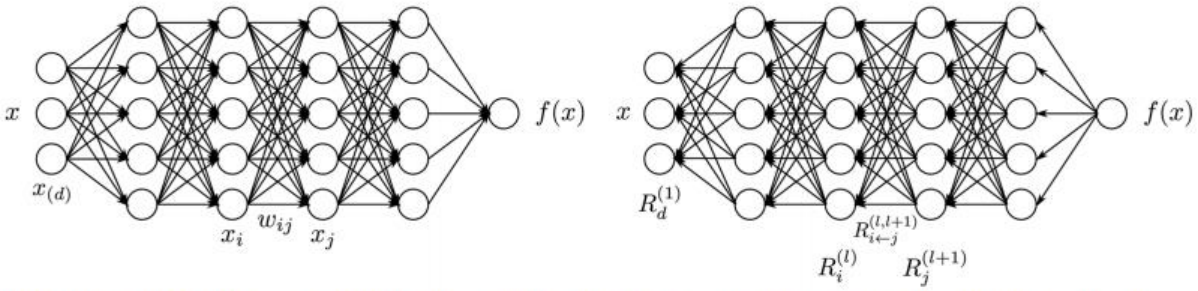

Fig 5. Multilayer neural network annotated with the different variables and indices describing neurons and weight connections. Left: forward pass. Right: backward pass.

doi:10.1371/journal.pone.0130140.g005

해당 방법은 다음과 같이 작동한다. 분류 결정 $f(x)$에 대한 특정 뉴런 $R_j^{(l+1)}$의 relevance를 알고, 이전 레이어의 뉴런에 전송된 메시지 $R_{i\leftarrow j}$의 관점에서 그러한 relevance의 decomposition을 얻고자 한다.

Relevance decomposition의 첫번째 가능한 선택은 Local과 global의 사전 활성화 비율에 근거하여 다음과 같이 주어진다.

$$R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{(l+1)} \qquad (19)$$

(19)의 propagation 규칙의 단점은 z_j의 값이 작을 때, relevance $R_{i\leftarrow j}$가 무한한 값을 가진다는 것이다. 이는 사전에 정의된 stabilizer $\varepsilon \geq 0$ 을 도입하여 극복할 수 있다. 특히, stabilizer가 매우 커지면 relevance는 완

전히 흡수된다.

$$R_{i\leftarrow j}^{(l,l+1)} = \begin{cases} \frac{z_{ij}}{z_j + \varepsilon} \cdot R_j^{(l+1)} & z_j \geq 0 \\ \frac{z_{ij}}{z_j - \varepsilon} \cdot R_j^{(l+1)} & z_j < 0 \end{cases} \qquad (20)$$

Relevance가 완전히 흡수될 염려가 없는 또 다른 방법으로는 negative나 positive로 각각 나누어 사전에 활성화하는 것이다. 이 방법은 다른 요인 α, β를 선택함으로써 positive와 negative의 중요도를 수동으로 제어할 수 있다.

$$\sum_i R_{i\leftarrow j}^{(l,l+1)} = R_j^{(l+1)} \cdot \left(1 - \frac{b_j^+}{2z_j^+} - \frac{b_j^-}{2z_j^-}\right) \qquad (21)$$

Discussion

비선형 learning machine은 현대 사회의 문제 해결에 있어 어디서나 볼 수 있다. Hard classification, regression, ranking problem 등의 다양한 분야에서 매우 성공적이었지만, 지금까지 해당 모델이 지닌 비선형성으로 인해 예측 결과에 대한 추가적인 설명이 없어 해결된 문제의 본질에 대한 더 나은 이해는 어려웠다. 그래서 우리는 pixel측면에서 비선형 이미지 분류 결정을 분해하여 사용자를 위한 transparency을 증진시키기 위한 방법을 도입했다. 즉, 예측된 클래스와 관련된 픽셀들을 강조하는 heatmap을 생성한다.

실제로 우리는 pixel-wise decomposition에 대해 두 가지 다른 접근법을 제안한다. 첫번째 taylor-type decomposition은 예측된 클래스의 가장 관련이 적은 데이터 포인트 근처에서 taylor-type decomposition을 수행함으로써 relevance score를 선형적으로 근사하는 방식이다. 두번째 layer-wise relevance propagation은 주어진 레이어에서 발견되는 클래스의 relevance를 이전 레이어에 분배하는 propagation 규칙을 적용하는 방식이다.

우리는 실험을 통해 BoW, neural network와 같은 비선형 분류기 모델에 taylor-type decomposition, layer-wise relevance propagation을 적용하면 학습된 분류기의 여러 측면에서 반영한 정보에 입각한 heatmap을 생성함을 보여주었다.

Fig 7. Pixel-wise decomposition for Bag of Words features over a histogram intersection kernel using the layer-wise relevance propagation for all subsequent layers and rank-mapping for mapping local features. Each triplet of images shows—from left to right—the original image, the pixel-wise predictions superimposed with prominent edges from the input image and the original image superimposed with binarized pixel-wise predictions. The decompositions were computed on the whole image. Images twice by Pixabay users tpsdave, and by Pixabay users sirocumo and Pixeleye.

doi:10.1371/journal.pone.0130140.g007

마지막으로 간단한 시각적 평가를 넘어 heatmap의 정보가 잘 반영된 정도를 평가하는 방법은 아직 명확하지 않다. 하지만 본 논문은 이와 유사한 품질을 가지는 heatmapping 방법을 구별할 수 있는 pixel-flipping 방법을 출발점으로 제안한다.