ReLU 신경망의 경로적 설명

임성우 [O1,4] 이주형 [2,3] 박치연 [3] 최재식 [4]

[1] 울산과학기술원 컴퓨터공학부

[2] 애리조나 주립대학 컴퓨팅, 정보 및 의사결정 시스템 공학부

[3] 삼성 리서치

[4] 카이스트 김재철 AI 대학원

seongwoolim@kaist.ac.kr, joolee@asu.edu, chiyoun.park@samsung.com, jaesik.choi@kaist.ac.kr

Pathwise Explanation of ReLU Neural Networks

Seongwoo Lim[O1,4] Joohyung Lee[2,3], Chiyoun Park[3], Jaesik Choi[4]

[1]Ulsan National Institute of Science and Technology

[2]Arizona State university

[3]Samsung Research

[4]Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology

요 약

인공지능이 어떻게 의사결정을 내렸는지 설명하는 기술(eXplainable AI, XAI)은 현실 상황에 인공지능을 신뢰성 있게 적용하기 위해 필수적이다. 최근 시각, 언어 등 다양한 분야에서 우수한 성능을 달성한 깊은 신경망(Deep Neural Network) 기반 모델은 의사결정 과정을 사람이 이해하기 어렵다는 문제가 존재한다. 본 논문에서는 ReLU 활성함수를 사용하는 신경망 모델의 국소적 선형 특징(Locally Linear Attribute)에 기반하여 경로적 설명 방법을 제시한다.

1. 서 론

깊은 신경망(Deep Neural Network) 기반 인공지능 모델들은 최근 시각, 언어, 제어 등 다양한 분야에서 우수한 성능을 달성했다. 하지만 깊은 신경망 기반 모델은 의사결정 과정을 사람이 이해하기 어렵기 때문에, 현실 상황에 적용하기 어려운 문제가 존재한다. 그러므로 인공지능 모델이 어떻게 의사결정을 내렸는지 설명하는 기술(explainable AI, XAI)은 현실 상황에 인공지능을 신뢰성 있게 적용하기 위해 필수적이다[1].

기존 XAI 기술은 입력과 출력 사이의 변화도(gradient)를 계산하여 설명하는 기술[2], 입력에 임의의 변화(perturbation)를 주었을 때 생기는 출력의 변화를 이용해서 설명하는 기술[3], 은닉 뉴런의 위치 정보를 유지하는 설명가능한 모델[4] 등 다양한 방법이 제시되었다.

본 논문에서는 최근 깊은 신경망에서 주로 사용되는 ReLU (Rectified Linear Unit) 활성함수 기반 신경망을 설명하는 것을 목표로 한다. ReLU 활성함수 기반 신경망은 주어진 입력에 대하여 신경망이 선형(Linear) 모델로 표현된다는 특징(Locally Linear Attribute)을 가진다[5]. 본 연구에서는 특정 입력에 대해 표현된 선형 모델이 다른 대부분의 입력에 대해 비슷한 출력만을 생성(biased)하는 현상을 발견했다. (그림 1) 즉, 표현된 선형 모델은 가중치(weight)보다는

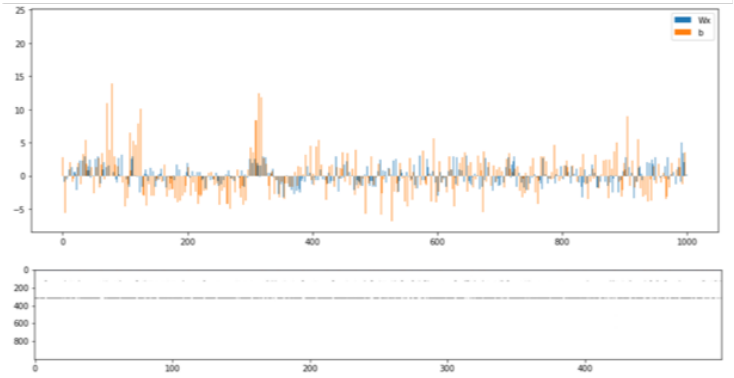

그림 1 (위) 특정 입력에 대해 표현된 선형 모델의 가중치와 입력의 곱(파란색)과 편향(주황색) (아래) 표현된 선형 모델의 다른 입력에 대한 출력 결과

편향(bias)에 더 의존한다. 이는 표현된 선형 모델의 가중치 기반 설명의 설명성이 낮을 수 있다는 것을 의미한다.

본 논문에서는 ReLU 신경망의 입력부터 출력까지의 각 경로별 선형 모델을 통해 설명을 생성하는 방법을 제안한다. 생성된 선형 모델은 ReLU 신경망을 설명하는 최소단위의 선형 모델을 표현 가능하다.

2. 시스템 정의

본 논문에서는 선형(합성곱, 선형 등) 혹은 단편적

선형(ReLU, 풀링 등)인 함수만을 포함한 신경망을 목표로 한다.

1) 주어진 d_0 차원 입력 $x \in \mathbb{R}^{d_0}$ 에 대해 N 개의 계층을 갖는 ReLU 신경망 $f: \mathbb{R}^{d_0} \to \mathbb{R}^{d_{N+1}}$ 은 d_{N+1} 차원 출력 $f(x)$를 출력한다.
2) $h^{(i)} = [h_1^{(i)}, \dots, h_{d_i}^{(i)}]$는 i번째 계층 $layer_i$에 대해서, **ReLU 활성함수를 적용하기 이전**의 d_i 차원 은닉 뉴런 벡터다.
3) $W^{(i)} \in \mathbb{R}^{d_i \times d_{i-1}}$ 는 $layer_{i-1}$ 과 $layer_i$ 사이의 가중치다.
4) $W_{j,k}^{(i)}$ 는 $layer_{i-1}$ 의 j번째 뉴런과 $layer_i$ 의 k번째 뉴런 사이의 가중치다.
5) $W_{j,*}^{(i)}$ 는 $layer_{i-1}$ 의 j번째 뉴런과 $layer_i$ 의 모든 뉴런 사이의 가중치다.
6) $b^{(i)}$는 $layer_i$의 편향이다.

3. ReLU 신경망의 경로적 설명

ReLU 함수는 0보다 큰 입력에 대해서는 입력값을 출력하고, 그 외의 입력에 대해서는 0을 출력하는 함수다. 이는 다음과 같이 정의 할 수 있다.

$$ReLU(x) = x \cdot \phi(x)$$

$\phi(x)$는 0보다 큰 입력에 대해서는 1을 출력하고, 그 외의 입력에 대해서는 0을 출력하는 함수다.

위에서 정의한 ReLU 함수와 N개의 선형 계층을 갖는 신경망 f는 다음과 같이 출력을 계산한다.

$$h^{(1)} = W^{(1)}x + b^{(1)}$$

$$h^{(2)} = W^{(2)}\left(h^{(1)}\phi\left(h^{(1)}\right)\right) + b^{(2)}$$

...

$$f(x) = W^{(N+1)}\left(h^{(N)}\phi\left(h^{(N)}\right)\right) + b^{(N+1)}$$

각 계층의 첫번째 뉴런의 집합으로 정의된 경로 $path_1 = \left[h_1^{(1)}, h_2^{(2)}, \dots, h_1^{(N)}\right]$의 선형 모델 표현 $f^{path_1}(x)$은 위의 수식에서 $\prod_{h \in path_1} \phi(h)$ 를 갖는 모든 항으로 정의한다.

$$f^{path_1}(x) = W_{1,*}^{(N+1)} W_{1,1}^{(N)} \dots W_{1,1}^{(2)}\left(W_{*,1}^{(1)}x + b^{(1)}\right)\prod_{i=1}^{N}\phi\left(h^{(i)}\right)$$

만약 경로 $path_1$ 의 뉴런 중 하나의 뉴런이라도 음수의 값의 가진다면, 정의된 선형 모델은 0의 값을 갖는다. 그러므로, 정의된 선형 모델 표현은 주어진 경로의 뉴런들이 유의미한 값(양수)를 가질 경우에만 ReLU 신경망에서 사용된다.

$f^{path_1}(x)$ 의 가중치 W^{path_1} 와 편향 b^{path_1} 은 다음과 같이 표현 가능하다.

$$W^{path_1} = W_{1,*}^{(N+1)} W_{1,1}^{(N)} \dots W_{1,1}^{(2)} W_{*,1}^{(1)}$$

$$= \frac{df(x)}{dh_1^{(N)}} \frac{dh_1^{(N)}}{dh_1^{(N-1)}} \dots \frac{dh_1^{(2)}}{dh_1^{(1)}} \frac{dh_1^{(1)}}{dx}$$

$$b^{path_1} = W_{1,*}^{(N+1)} W_{1,1}^{(N)} \dots W_{1,1}^{(2)} b^{(1)} = \frac{df(x)}{dh_1^{(N)}} \frac{dh_1^{(N)}}{dh_1^{(N-1)}} \dots \frac{dh_1^{(2)}}{dh_1^{(1)}} b^{(1)}$$

$$= \frac{df(x)}{dh_1^{(N)}} \frac{dh_1^{(N)}}{dh_1^{(N-1)}} \dots \frac{dh_1^{(2)}}{dh_1^{(1)}} \left(h_1^{(1)} - \frac{dh_1^{(1)}}{dx} x\right)$$

경로는 반드시 모든 계층의 뉴런을 포함할 필요가 없으며, 한 계층에서 여러 뉴런을 포함할 수도 있다. 극단적인 예로, $path_2 = \left[h_k^{(n)}\right]$ 가 가능하다. 해당 경우 선형 모델 $f^{path_2}(x)$ 의 가중치 W^{path_2} 와 편향 b^{path_2} 는 다음과 같이 표현한다.

$$W^{path_2} = \frac{df(x)}{dh_k^{(n)}} \frac{dh_k^{(n)}}{dx}$$

$$b^{path_2} = \frac{df(x)}{dh_k^{(n)}} \left(h_k^{(n)} - \frac{dh_k^{(n)}}{dx} x\right)$$

4. 경로 생성 알고리즘

Algorithm 1 Hierarchical decomposition

Input: h: set of all hidden units in N layer ReLU NN,
        y: NN output, $index$: target index

$path = \{\}$
$W^{(N+1)} = \frac{d}{dh^{(N)}} y$
$Importance = elemenwiseMultiply(W_{*,index}^{(N+1)}, h^{(N)})$
$index = argmax(Importance)$
$path.add(h_{index}^{(N)})$
for $n \leftarrow N-1$ to 0 **do**
    $W^{(n+1)} = \frac{d}{dh^{(n)}} h^{(n+1)}$
    $Importance = elemenwiseMultiply(W_{*,index}^{(n+1)}, h^{(n)})$
    $index = argmax(Importance)$
    $path.add(h_{index}^{(n)})$
return $path$

깊은 신경망 모델에 대해서 모든 가능한 경로를 다루는 것은 계산상 어려운 문제이기 때문에, 하향식(Top-down) 계층적 분해(Hierarchical decomposition) 알고리즘[6]을 응용한다.

계층적 분해 알고리즘(Algorithm 1)은 상위 계층(upper layer)의 목표 뉴런에 대해 가장 큰 영향력을 갖는 뉴런을 찾는 것을 목표로 한다. 영향력을 계산할 때는 가중치와 뉴런값의 곱으로 표현되며(6째줄),

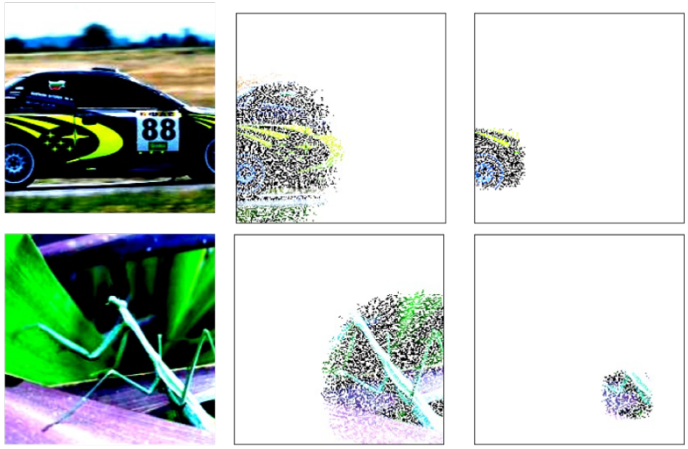

그림 3 계층적 분해 알고리즘으로 생성된 경로의 선형 모델 가중치 시각화 (좌) 입력 이미지 (중) 얕은 경로의 시각화 (우) 깊은 경로의 시각화

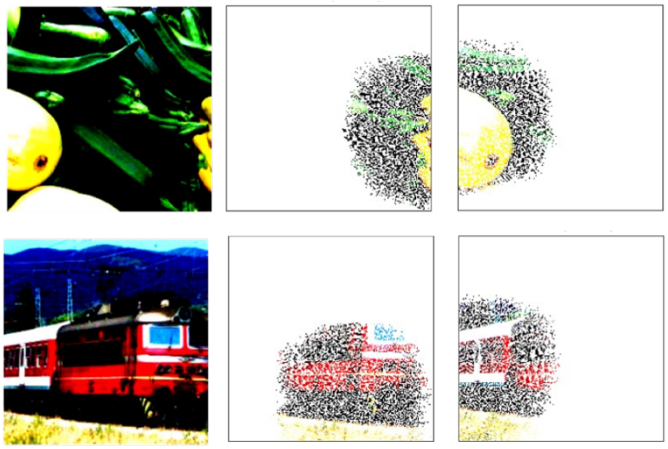

그림 2 그럴듯한(plausible) 카테고리에 대한 설명 시각화 (좌) 입력 이미지 (중) 목표 카테고리에 대한 설명 시각화 (우) 그럴듯한 카테고리에 대한 설명 시각화

가장 큰 영향력을 갖는 뉴런이 경로에 추가된다(8째줄). 경로에 뉴런이 추가된 후, 추가된 뉴런이 새로운 목표 뉴런이 되며 같은 작업을 하위 계층에서 반복한다(9-13째줄).

5. 경로적 설명의 시각화

이 장에서는 ImageNet 데이터[7]에 학습된 VGG-16[8] 모델의 경로적 설명을 시각화를 통해 보여준다. ImageNet 데이터는 1000 종류의 출력 카테고리가 있지만, 본 논문에서는 임의의 10 종류의 출력 카테고리에 대해서 상위 선형 모델을 미세 조정(fine-tuning)하여 실험을 진행하였다.

그림 2는 3장의 경로 생성 알고리즘으로 생성된 경로의 선형 모델 가중치 시각화 결과이다. 포함된 뉴런의 개수가 적은 얕은 경로에 대해서는 더욱 포괄적인 정보가 시각화 되고(그림 2의 (중)), 뉴런의 개수가 많은 깊은 경로 일수록 국소적인 정보가 시각화 된다(그림 2의 (우)). 이와 같은 현상은 얕은 경로가 깊은 경로들의 합으로 표현이 가능하기 때문에 나타난다.

그림 3은 목표 카테고리를 제외한 나머지 카테고리 중 가장 높은 출력값을 갖는(plausible) 카테고리에 대한 시각화 결과이다. 첫번째의 실제 주키니 호박 카테고리를 갖는 입력에 대해, 피망 카테고리라고 판단할 때 사용되는 입력 특징(그림 3의 (우))을 보여준다. 두번째는 실제 전기 기관차 카테고리를 갖는 입력에 대해, 감옥 카테고리라고 판단할 때 사용되는 입력 특징을 보여준다. 이와 같은 분석을 통해 모델이 잘못된 판단을 내렸을 경우, 어떤 의사결정 과정으로 잘못된 판단을 내렸는지를 확인할 수 있고, 어떻게 하면 정확한 판단을 내릴 수 있는지 개선이 가능하다.

6. 결론 및 향후 연구

본 연구에서는 ReLU 활성함수를 사용하는 신경망 모델의 경로적 설명 방법을 제안하였다. 3장에서 어떻게 ReLU 신경망이 경로적으로 표현가능한지와 각 경로에 대응되는 선형 모델을 계산하는 과정을 보였고, 4장에서 경로를 하향식 계층적 분해 알고리즘으로 목표 출력을 최대화하는 경로를 찾는 방법을 제안하였다. 5장에서는 생성된 경로에 대응되는 선형 모델의 시각화 및 실제 카테고리가 아닌 임의의 카테고리에 대해서도 설명이 가능함을 보여 신경망 모델 의사결정 과정의 설명 뿐만 아니라 개선에 도움이 되는 설명을 생성해 보였다.

ReLU 활성함수 뿐만 아니라 $x \cdot \phi(x)$ 의 형식으로 표현될 수 있는 다양한 활성함수에도 경로적 설명 방법을 적용하고, 보다 효율적이고 의미있는 경로 생성 알고리즘을 개발하는 것이 앞으로 연구할 과제이다. 또한, 가중치 뿐만 아니라 편향에 대한 설명제공 또한 앞으로 연구할 과제이다.

참 고 문 헌

[1] David Gunning et al. XAI - explainable artificial intelligence. Sci. Robotics, 4(37), 2019

[2] Karen Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR, 2014.

[3] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. ECCV, 2014.

[4] Ramprasaath R. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. ICCV, 2017.

[5] Benjamin Sattelberg et al. Locally linear attributes of relu neural networks. CoRR, 2020.

[6] Ming-Ming Cheng et al. Deeply explain CNN via hierarchical decomposition. Int. J. Comput. Vis.,

131(5):1091-1105, 2023.
[7] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis., 115(3):211-252, 2015.
[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. ICLR, 2015.