그래프 합성곱 신경망에 대한 기울기(Gradient) 기반 설명 기법

김채현, 이기용
숙명여자대학교 컴퓨터과학과
email : {7chaeny25, kiyonglee}@sookmyung.ac.kr

A Gradient-Based Explanation Method for Graph Convolutional Neural Networks

Chaehyeon Kim, Ki Yong Lee
Department of Computer Science, Sookmyung Women's University

요 약

설명가능한 인공지능은 딥러닝과 같은 복잡한 모델에서 어떠한 원리로 해당 결과를 도출해냈는지에 대한 설명을 함으로써 구축된 모델을 이해할 수 있도록 설명하는 기술이다. 최근 여러 분야에서 그래프 형태의 데이터들이 생성되고 있으며, 이들에 대한 분류를 위해 다양한 그래프 신경망들이 사용되고 있다. 본 논문에서는 대표적인 그래프 신경망인 그래프 합성곱 신경망(graph convolutional network, GCN)에 대한 설명 기법을 제안한다. 제안 기법은 주어진 그래프의 각 노드를 GCN을 사용하여 분류했을 때, 각 노드의 어떤 특징들이 분류에 가장 큰 영향을 미쳤는지를 수치로 알려준다. 제안 기법은 최종 분류 결과에 영향을 미친 요소들을 gradient를 통해 단계적으로 추적함으로써 각 노드의 어떤 특징들이 분류에 중요한 역할을 했는지 파악한다. 가상 데이터를 통한 실험을 통해 제안 방법은 분류에 가장 큰 영향을 주는 노드들의 특징들을 실제로 정확히 찾아냄을 확인하였다.

1. 서론

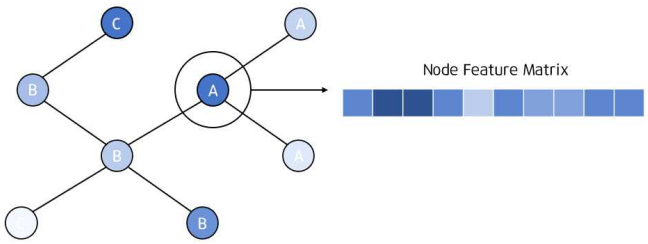
설명가능한 인공지능(eXplainable artificial intelligence, XAI)이란 딥러닝 모델 내부를 분석하여 판단 결과가 도출된 이유를 사용자가 이해하고 신뢰할 수 있도록 설명하는 기술이다. 딥러닝 모델은 다양한 분야에서 좋은 성능을 내고 있지만, 모델의 복잡성으로 인해 사람이 해당 모델을 이해하고 신뢰하기는 어렵게 만든다. 뿐만 아니라 모델이 예측한 결과가 신뢰할만한지, 모델이 공정한 의사결정을 내렸는지 등의 여부를 파악하기도 어렵다. 하지만 의료 인공지능과 같이 모델을 신뢰할 수 있는지 정확히 판단해야 하는 분야들이 존재한다. 가령, 의료분야에서는 모델이 잘못 학습된다면 의료사고까지 이어질 수 있기 때문에 딥러닝 모델의 경우 임상 적용에 한계가 있다. 따라서 최근 모델의 신뢰성과 안정성을 보장하기 위해 딥러닝 모델에서 판단 결과를 뒷받침하는 근거를 인식하는 XAI를 적용하는 연구가 활발히 진행되고 있다[1].

XAI의 기법은 크게 ABM(activation-based methods), BBM(backpropagation-based methods), 그리고 PBM(perturbation-based methods)로 나뉜다[2][3][4]. ABM은 이미지 분류 모델에서 입력 데이터의 어떤 부분이 클래스를 예측하는 데 영향을 주었는지 시각화하는 방법이며, BBM은 딥러닝 모델에서 예측 결과를 역추적하여 신경망의 각 계층별 기여도를 측정하는 방법이다. 마지막으로 PBM은 입력에 노이즈(noise)를 추가하여 모델의 결과를 분석하는 방법이다.

하지만 지금까지 대부분의 XAI 연구는 이미지 처리 딥러닝 모델에서 이루어졌다. 따라서 기존 방법들을 단순히 그래프 데이터에 적용한다면, 그래프 데이터의 중요한 특징인 데이터 간의 관계를 반영하지 못하는 문제가 발생한다. 그래프 데이터는 각 노드에 대한 특징(feature) 정보를 담고 있는 특징행렬(node feature matrix)과 노드 사이의 연결성을 나타내는 인접행렬(adjacency matrix)로 표현된다. 이와 같이 하나의 행렬로 나타내어지는 이미지 데이터와는 다르게, 그래프 데이터의 경우는 두 개의 행렬을 사용하여 각 데이터의 특징과 데이터 간의 관계를 표현한다.

따라서 본 논문에서는 그래프 데이터의 분류에 가장 널리 사용되는 그래프 신경망인 그래프 합성곱

신경망(graph convolutional network, GCN)[5]에 대한 설명 기법을 제안한다. 제안 방법은 특징행렬과 인접행렬에 대한 정보를 모두 활용하여 GCN의 분류 결과에 가장 큰 영향을 준 특징들을 파악한다. 제안 방법은 주어진 그래프의 각 노드의 레이블(label)을 GCN으로 분류했을 때, 각 노드의 어떤 특징들이 분류에 가장 큰 영향을 미쳤는지를 수치로 알려준다. 제안 기법은 최종 분류 결과에 영향을 미친 요소들을 각 층(layer)마다 기울기(gradient)의 값을 구하여 단계적으로 역추적한다. 역추적이 완료되면 각 노드의 어떤 특징들이 분류에 중요한 역할을 했는지가 가중치 형태로 출력된다. 본 논문에서는 가상 데이터에 대한 간단한 실험을 통해 제안 방법이 실제로 노드의 분류에 가장 큰 영향을 주는 특징들을 정확히 찾아냄을 확인하였다. (그림 1)은 제안 방법의 용도를 나타내는 개념도이다. 각 노드가 A, B, C로 분류되었을 때, 각 노드의 특징들 중 어떤 특징들이 분류에 가장 큰 영향을 미쳤는지를 수치로 보여주며, 이 수치들을 시각화에 사용할 수 있다.



(그림 1) 제안 방법의 개념도

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 간략히 설명하고, 3장에서는 본 논문의 제안 방법을 구체적으로 설명한다. 4장에서는 가상 데이터에 기반한 실험 결과를 제시하고, 마지막 5장에서 결론을 맺는다.

2. 관련 연구

지금까지 제안된 대표적인 XAI 기법들에는 LIME(local interpretable model-agnostic explanations)[2], Grad-CAM(gradient-class activation map)[3], 그리고 LRP(layer-wise relevance propagation)[4]가 있다. LIME은 각각의 데이터에 대한 해석과정을 보여주는 알고리즘이다. LIME은 텍스트 혹은 이미지 데이터와 같이 행렬로 표현되는 데이터에 작동하는 XAI 기법으로, 입력 이미지 데이터를 해석 가능하도록 인식 단위로 분해하여 해석한다. 그 후, 나누어진 영역을 조합하여 모델이 대상을 가장 잘 분류할 수 있는 대표 이미지를 구성한다. Grad-CAM은 CNN(convolutional neural network) 모델에서 마지막 층으로 전달되는 미분값의 정보를 이용하여 의사결정과 관련된 각 뉴런의 중요도를 이해하는 방법이다. Grad-CAM은 합성곱층에서 역전파(backpropagation) 과정에서의 미분값을 활용하여 해당 미분값과 합성곱층의 특징 맵(feature map)을 곱하고 추가적으로 ReLU 활성화함수를 거친 값을 입력 이미지에 히트맵(heatmap) 형태로 표현한다. LRP는 모델의 결과에 대한 기여도를 출력에서 입력 방향으로 역추적해서 입력 이미지에 히트맵을 출력한다. 특징 결과가 나오게 된 원인을 분해하고 기여도를 재분배하는 타당성 전파 기법과 타당성 전파를 통해 얻어낸 원인을 가중치로 설정하고 각 특징이 결과에 얼마나 영향을 미치는지 해체하는 분해 기법을 혼합하여 사용한다.

한편 그래프 데이터는 소셜 네트워크나 단백질 구조와 같이 실생활에서 다양한 정보를 표현하는 데이터이다. 최근 그래프 신경망(graph neural network)에 대한 연구가 활발히 진행되면서 그래프 또는 그래프 내의 각 노드를 신경망을 사용하여 분류하는 기술이 활발히 연구되고 있다. 전통적인 XAI 기법들은 주로 CNN에 대해 이루어졌다. 하지만 그래프 신경망의 경우 그래프 구조까지 고려해야 하므로 전통적인 XAI 기법을 그대로 적용하기 어려우며, 따라서 최근 그래프 데이터만을 위한 XAI 연구들이 시도되고 있다. 특히 대표적인 그래프 신경망인 GCN에 대해서는 아직까지는 제한적인 설명 기법만 제안되었으며[6], 그에 따라 본 논문에서는 GCN의 분류 결과에 대한 설명 기법을 제안한다.

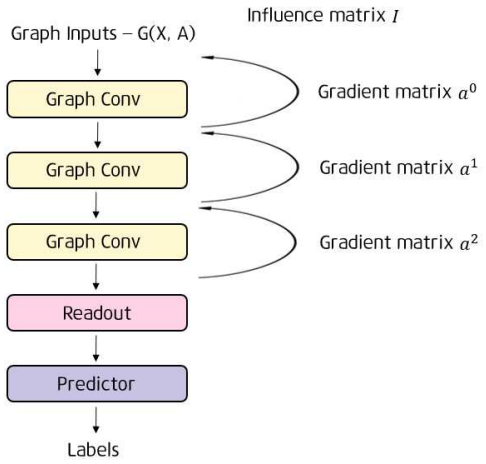
3. 제안 방법

본 논문에서는 주어진 그래프의 각 노드의 레이블을 GCN으로 분류할 때, 분류 결과에 가장 큰 영향을 미친 노드의 특징들을 탐색하는 방법을 제안한다. 그림 2는 제안하는 기법의 흐름을 나타낸다. 제안 방법은 (1) 노드 분류 단계와 (2) 노드 분류 결과 설명 단계로 구성되며, 노드 분류 결과 설명 단계는 GCN을 역으로 추적하여 노드 분류 결과에 가장 큰 영향을 미친 각 노드의 특징들을 탐색한다.

3.1 노드 분류 단계

GCN은 이미지에 대한 합성곱을 그래프 데이터로 확장한 인공신경망 모델이다. GCN 모델의 입력

은 각 노드의 특징들을 나타내는 $N \times F$ 크기의 특징행렬 X와, 노드 간의 연결상태를 나타내는 $N \times N$ 크기의 인접행렬 A로 구성된다. 단, N은 노드의 수를, F는 노드 특징들의 개수를 나타낸다.



(그림 2) 제안 방법의 흐름도

GCN을 사용하여 주어진 그래프의 노드들을 분류하는 과정은 다음과 같다. 주어진 그래프의 정보를 나타내는 행렬 X와 A가 주어지면 여러 개의 그래프 합성곱층(graph convolutional layer)를 거쳐 각 노드의 내재된 특징들이 추출된다. 각 그래프 합성곱층은 각 노드에 대해 그와 연결되어 있는 노드들의 정보를 모아 해당 노드의 특징행렬의 값을 갱신한다. 예를 들어 첫 번째 그래프 합성곱층은 특징행렬과 인접행렬을 한 번 곱함으로써, 각 노드에 대해 그와 연결된 노드의 특징을 모두 더한 값으로 해당 노드의 특징을 갱신한다. 제안 모델에서는 크기가 1×1이고, 스트라이드(stride)가 1인 필터를 사용하여 특징을 추출하였으며, 특징행렬이 더 고차원적인 정보를 담을 수 있도록 3번의 그래프 합성곱층을 수행하였다. 다음은 각 그래프 합성곱층에서 이루어지는 연산을 나타내는 식이다.

$$H^{(l+1)} = \sigma(AH^{(l)}W + B)$$

$H^{(l)}$는 l번째 층에서 얻어진 특징행렬, A는 인접행렬, W와 B는 각각 학습으로 얻어진 가중치(weight)와 편향(bias) 행렬, 그리고 σ는 ReLU 활성화 함수를 의미한다. 합성곱층을 모두 거쳐 얻어진 최종 특징행렬은 판독층(readout layer)를 거쳐 하나의 벡터 형태로 변환된다. 마지막으로 예측층(predictor layer)은 판독층의 출력 벡터를 받아 각 노드마다 그의 분류된 레이블을 출력한다.

3.2 노드 분류 결과 설명 단계

3.1절에서 GCN 모델이 주어진 그래프의 노드들을 분류하고 나면, 본 단계에서는 분류 결과에 영향을 준 주요 특징들을 역추적한다.

제안 방법은 각 그래프 합성곱층에서 특징행렬 $H^{(l)}$의 각 원소 $H^{(l)}(i,j)$가 다음 특징행렬 $H^{(l+1)}$에 미치는 영향력을 다음 식과 같이 $H^{(l)}(i,j)$의 값에 대한 $H^{(l)}$의 기울기(gradient)로 나타낸다.

$$a^{(l)}(i,j) = \frac{\partial H^{(l+1)}}{\partial H^{(l)}(i,j)}$$

즉, $a^{(l)}(i,j)$는 $H^{(l)}(i,j)$가 $H^{(l+1)}$에 미치는 영향력을 $H^{(l)}(i,j)$ 값의 증가에 따른 $H^{(l+1)}$ 값의 증가율, 즉 기울기로 나타내는 값이며, $a^{(l)}$는 $H^{(l)}$의 각 원소 $H^{(l)}(i,j)$가 다음 특징행렬 $H^{(l+1)}$에 미치는 영향력을 나타내는 행렬이 된다.

또한 (그림 2)에서 나타낸 바와 같이 그래프 합성곱층은 여러 개가 있을 수 있으므로, 각 합성곱층마다 $a^{(l)}$를 구한 후 $a^{(0)}, a^{(1)}, a^{(2)}, \ldots$의 값들을 원소별(element-wise)로 곱하여 모든 그래프 합성곱층을 반영한 최종 기울기 행렬을 구한다. 마지막으로 이를 주어진 그래프의 특징행렬 X과 다시 원소별로 곱하여 특징행렬의 각 원소 $X(i,j)$가 최종 분류 결과에 미친 영향력을 최종적으로 얻는다. 다음은 지금까지 설명한 식을 나타낸다.

$$I(i,j) = X(i,j) \circ a^{(0)}(i,j) \circ a^{(1)}(i,j) \circ a^{(2)}(i,j)$$

위 수식에서 $X(i,j)$는 입력 특징행렬 X의 각 원소를 의미하며, $a^{(0)}(i,j)$는 $X(i,j)$가 첫 번째 합성곱층을 통해 $H^{(1)}$에 미치는 영향력, $a^{(1)}(i,j)$는 $H^{(1)}(i,j)$가 두 번째 합성곱층을 통해 $H^{(2)}$에 미치는 영향력, $a^{(2)}(i,j)$는 $H^{(2)}(i,j)$가 세 번째 합성곱층을 통해 $H^{(3)}$에 미치는 영향력을 각각 나타낸다. $\circ$는 아다마르 곱(Hadamard product)을 나타낸다.

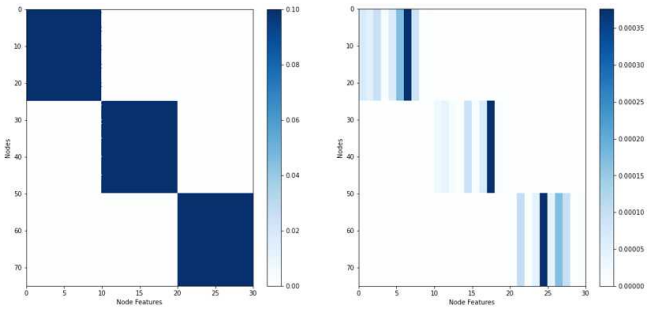
이러한 방법을 사용하면 최종적으로 $I(i,j)$를 통해 입력 특징행렬 X의 원소 $X(i,j)$가 GCN을 통해 최종 분류 결과에 미친 영향력을 상대적인 수치로 얻을 수 있다.

4. 실험 결과

본 장에서는 제안 방법이 GCN이 노드 분류에 영향을 미치는 노드 특징들을 올바르게 탐색하는지

여부를 실험하였다.

실험에서는 75개의 노드를 가지는 그래프를 가정하였으며, 각 노드는 30개의 특징을 가진다고 가정하였다. 따라서 특징행렬 X는 75×30의 크기를 가지고 인접행렬 A는 75×75의 크기를 가진다. 노드 분류의 영향을 미치는 특징들을 의도적으로 만들기 위해 가상 데이터를 생성하였다. 즉, 특징행렬에서 1번째부터 10번째 특징들의 값이 모두 1인 노드들(노드 0~25번)은 A 클래스로, 11번째부터 20번째 특징들의 값이 모두 1인 노드들(노드 26~50번)은 B 클래스로, 그리고 21번째부터 30번째 특징들의 값이 모두 1인 노드들(노드 51~75번)은 C 클래스로 레이블을 주어 GCN을 훈련시켰다. 또한 그래프의 구조는 모든 노드 간에 연결선이 존재하는 완전 연결 그래프를 가정하였다. 실험에 사용된 GCN과 제안 방법은 모두 PyTorch로 구현하였으며, GCN 훈련 및 제안 방법의 실행은 RTX 2080 Super GPU와 Intel i7-9700 3.3GHz CPU, 16GB RAM, 1TB HDD가 탑재된 PC에서 수행하였다.



(a) 입력 특성행렬 X (b) 영향력 행렬 I

(그림 3) 제안 방법의 실험 결과

그림 3(a)는 가상으로 생성한 입력 특징행렬 X을 시각화한 그림이다. 값이 1인 특성은 진하게, 값이 0인 특성은 연하게 나타내었으며, 노드 1~25번은 A 클래스, 노드 26~50번은 B 클래스, 그리고 노드 51~75번 C 클래스로 분류되어야 한다. 그림 3(b)는 제안 방법을 사용하여 X의 각 원소가 분류 결과에 미치는 영향력을 나타내는 영향력 행렬 I를 구하고 이를 시각화한 그림이다. 그림 3(a)와 유사하게 값이 큰 원소들은 진하게, 값이 작은 원소들은 연하게 나타내었다.

그림 3(b)에서 볼 수 있듯이 제안 설명 기법으로 계산한 영향력 행렬 I는 원 특징행렬 X에서 분류 결과에 영향을 미치는 값들을 비교적 정확하게 탐색해냄을 확인할 수 있다. 즉, 노드 0~25번에 대해서는 1번째부터 10번째 특징들의 영향력이 높게 나왔으며, 노드 26~50번에 대해서는 11번째부터 12번째 특징들의 영향력이 높게 나왔고, 노드 51~75번에 대해서는 21번째부터 30번째 특징들의 영향력이 높게 나왔다. 따라서 제안 방법은 각 노드가 GCN으로 분류되었을 때, 어떤 특징이 그에 가장 큰 영향력을 미쳤는지를 효과적으로 잘 찾아냄을 확인할 수 있다.

5. 결론

본 논문에서는 GCN이 그래프의 각 노드를 분류했을 때, 해당 노드가 그렇게 분류된 가장 큰 원인이 무엇인지를 설명하는 기법을 제안하였다. 제안 방법은 GCN을 구성하는 각 그래프 합성곱층에 대해, 합성곱층의 입력 특징행렬의 각 원소가 합성곱층의 출력 특징행렬의 값들에 미치는 영향을 기울기(gradient)로 계산하고, 이들을 사용하여 원 특징행렬의 각 원소가 최종 분류 결과에 미치는 영향력을 계산한다. 실험을 통해 제안 방법은 그래프를 구성하는 각 노드의 클래스를 결정하는 특징들을 효과적으로 탐색함을 확인하였다.

Acknowledgement

이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2021R1A2C1012543).

참고문헌

[1] A. Arrieta, et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”, Information Fusion, vol. 58, pp. 82-115, Jun 2020

[2] M. T. Ribeiro, S. Singh and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier”, ACM SIGKDD, Aug 2016.

[3] R. R. Selvaraju, et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017

[4] S. Bach, et al., “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” PloS one, vol. 10, Jul 2015.

[5] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” International Conference on Learning Representations (ICLR), 2017.

[6] J. Hu, T. Li and S. Dong, “GCN-LRP explanation: exploring latent attention of graph convolutional networks,” International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2020.