설명 가능한 인공지능 기반 의료 AI 기술 연구 동향

김재현[1], 김유신[2], 이세종[3], 안세영[4], 노재원[5], 김종훈[6], 조성현*
한양대학교 인공지능융합학과[1], 한양대학교 컴퓨터공학과*[2,3,4,5,6]
바이오인공지능융합전공[1,2,3,4]
e-mail : {insam2802, hpwgg045, kingsaejong, tpdud1014, wodnjs1451, chopro, iproj2}@hanyang.ac.kr

A Survey on Artificial Intelligence based Explainable Artificial Intelligence

Jaehyeon Kim[1], Yushin Kim[2], Sejong Lee[3], Seyoung Ahn[4], Jaewon Noh[5], Jonghun Kim[6], Sunghyun Cho*
Dept. of Applied Artificial Intelligence, Hanyang University[1]
Dept. of Computer Science and Engineering, Hanyang University*[2,3,4,5,6]
Major in Bio-Artificial Intelligence[1,2,3,4]

Abstract

Artificial intelligence is being used in a variety of fields, including medicine. But their lack of interpretability and explainability stand as one of the main drawbacks. To solve this problem, explainable artificial intelligence has been studied in healthcare. In this paper, we summarize recent advances in explainable artificial intelligence technologies and introduce the use of explainable artificial intelligence studies in medicine.

I. 서론

인공지능 (Artificial intelligence, AI) 기술이 발전함에 따라 다양한 분야에서 활용되고 있다. 특히 의료 분야에서는 질병의 초기 발견 [1], 질환의 분류 [2], 최적화된 약물 용량 결정 [3] 등의 문제를 해결하려는 연구가 활발히 진행되고 있다. 하지만 AI 알고리즘의 복잡성으로 인해 블랙박스 문제가 발생한다. 블랙박스 문제 AI의 의사결정 근거 및 결정 도출 과정의 신뢰성을 보장할 수 없다. 이로 인해 강한 신뢰성을 필요로 하는 의료 분야의 특성상 AI의 의사결정을 전적으로 신뢰할 수 없다는 문제가 발생한다. 따라서 신뢰성을 보장하기 위해 의료 분야에 설명 가능한 AI (Explainable AI, XAI)를 적용하는 연구가 활발히 진행되고 있다. XAI 기술은 AI의 의사결정에 대해 해석적이고, 직관적이며 사람이 이해 가능한 설명을 제공한다. 이를 통해 AI의 의사결정에 대한 투명성 및 신뢰성을 확보할 수 있다.

II. 본론

본 장에서는 설명 가능한 AI의 대표적인 기법인 Class Activation Mapping (CAM) [4], Layer-wise Relevance Propagation (LRP) [5] 및 Local Interpretable Model Agnostic Explanations (LIME) [6]에 대하여 소개하고 해당 기술을 의료 분야에 활용한 연구에 대하여 기술한다.

2.1 CAM 기반 XAI 기술 활용 현황

CAM은 이미지 분류 모델에서 입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화하는 방법이다. CAM은 기존의 Convolution Neural Network (CNN)에서 사용하는 fully-connected layer 대신 global average pooling (GAP)을 사용한다. GAP는 convolution layer에서 분류하려는 클래스의 수만큼의 채널을 갖게 한 후 각 채널을 기준으로 평균 합을

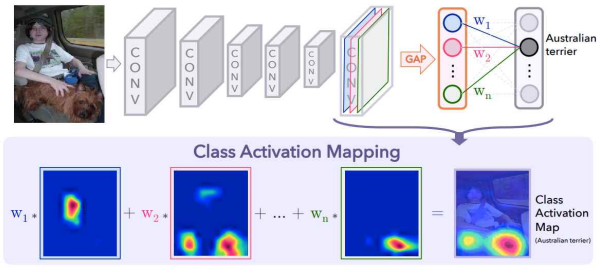

그림 1. CAM을 이용해 시각화된 heatmap

구한다. 이 각각의 값들은 클래스에 대응하는 값들이 되고 가장 큰 값을 가지는 부분으로 예측을 진행하게 된다. 또한 대응하는 값을 이용하여 그림 1과 같은 heatmap 형태로 시각화 하여 입력 데이터에서 가장 큰 영향을 준 부분을 직관적으로 이해할 수 있다. 따라서 이미지 분류를 위해 CAM 기반의 XAI 기술을 이용한 다양한 질병 분류 연구가 진행되었다.

[7]의 저자들은 자기 공명 영상(MRI) 이미지 데이터에서 근디스트로피를 분류하고 MRI 이미지의 중요한 부분을 heatmap 형태로 시각화 하였다. 해당 논문에선 CAM 기반의 Improve-CAM 기법을 이용하여 근디스트로피 이미지를 분류하였고 4겹 교차 검증을 통해 최대 91.7%의 평균 분류 성능을 보였다. 이는 딥러닝이 아닌 방법과 비교하여 약 40% 이상의 성능 향상을 보였고, ResNet 모델의 분류 성능 보다 약 5% 이상의 성능 향상을 보였다. 또한 MRI 이미지 데이터에서 정확한 눈 종양의 분류를 통해 안구암의 진단 및 치료를 진행하는 연구도 진행되었다. [8]의 저자들은 포도막 흑색종의 위치, 크기 및 외형과 관련된 특징을 얻기 위한 CAM기반 ResNet 모델과 포도막 흑색종의 분류를 위한 2D-Unet CNN의 조합으로 분류 모델을 설계하였다. 또한 안구 MRI 이미지에서 포도막 흑색종의 위치를 시각화 하였다. 해당 연구에서 제안하는 모델은 평균 84.5%의 분류 정확도를 보였다. 이는 3D-Unet, 3D-CNN 등의 기존 분류 모델과 비교하여 최대 약 20% 향상된 성능을 보인다.

2.2 LRP 기반 XAI 기술 활용 현황

LRP는 딥러닝 모델에서 예측 결과를 역추적하여 신경망의 각 계층별 기여도를 측정할 수 있는 방법이다. 이를 통해 그림 2와 같이 각 계층의 기여도를 heatmap 형태로 시각화 하여 직관적으로 이해할 수 있다. LRP는 타당성 전파와 분해 방법을 이용해 모델을 해부한다. 예측 결과로부터 타당성 점수를 입력단 방향으로 역추적 한 뒤 그 비중을 분배하여 계산한다. 신경망의 모든 뉴런들의 타당성 점수가 계산되면 마지막 계층의 타당성 점수를 heatmap 형태로 시각화 하

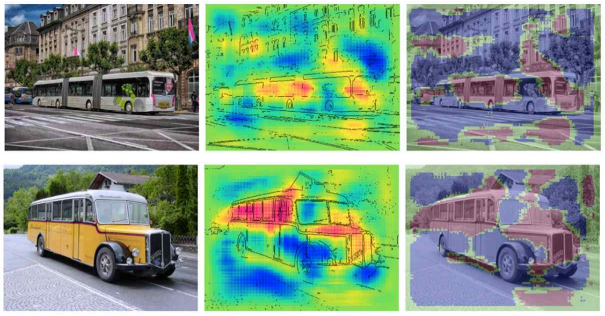

그림 2. LRP를 통해 시각화된 heatmap

여 중요도를 시각적으로 표현할 수 있다. 따라서 의료 분야에서 LRP 기법을 이용해 분류 모델의 타당성을 확인하기 위한 연구가 진행되고 있다.

[9]의 저자들은 Long Sort Term Memory (LSTM) 모델을 이용하여 호르몬 치료에 대한 실제 의료 데이터를 학습하고 LRP를 통해 시각화 하는 시뮬레이션을 진행하였다. 해당 모델은 81.2%의 정확도를 달성하였고 실제 임상의는 해당 모델이 내린 결정의 75.4%를 동의하는 것으로 나타났다. 또한 LRP를 적용한 각 수치들의 타당성 점수를 통해 항호르몬 요법과 화학 요법에서 어느 수치가 치료에 가장 큰 영향을 주는지를 시각화 하였다. 또한 3D-CNN 모델을 통해 다발성 경화증을 진단하는 연구도 진행되었다[10]. 해당 연구에서는 LRP를 통해 분류 모델의 의사결정을 설명 가능한 딥러닝 프레임워크를 제시한다. 제안하는 프레임워크는 평균 87. 04%의 분류 정확도를 달성하였다. 추가적으로 LRP를 통해 분류 모델의 성능이 병변의 유무에 큰 영향을 받지만 병변의 위치, 비병변 백질과 시상과 같은 회백질 영역과 같은 추가 정보를 사용한다는 사실을 보였다.

2.3 LIME 기반 XAI 기술 활용 현황

LIME은 모델이 데이터의 어느 영역을 분류의 근거로 사용했는지 설명하는 기법이다. LIME은 입력 데이터를 변형하여 인식 단위로 쪼개어 데이터를 해석한다. 분류에 영향을 많이 미치는 부분일수록 변형되었을 때 예측 결과가 크게 변하므로 입력을 변형시키면서 예측 결과를 크게 변화시키는 값을 찾는다. 또한 변형에 따른 예측 결과의 변화를 그림 3과 같은 방식으로 시각화 하여 중요도를 시각적으로 표현할 수 있

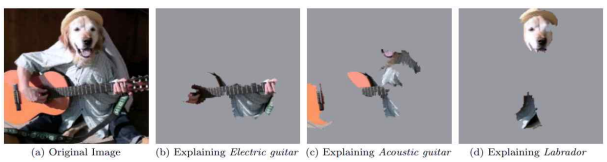

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

그림 3. LIME을 이용한 중요도에 따른 이미지 마스킹

표 1. 의료 분야에서 XAI 기술 활용 현황 비교

기법	확장성	연구	사용 분류 모델	분류 모델의 성능 수치 (%)	데이터 형식
CAM	CNN 모델	[7]	CNN	91.7 (Accuracy)	MRI Image
		[8]	ResNet	84.5 (Accuracy)	MRI Image
		[13]	CNN-CAM	98.5 (Accuracy)	Image
LRP	딥러닝 모델	[9]	LSTM	81.2 (Accuracy)	Numeric
		[10]	CNN	88.4 (Accuracy)	MRI Image
		[14]	CNN	88.0 (Accuracy)	MRI Image
		[15]	CNN	86.7 (Accuracy)	MRI Image
LIME	모든 모델	[11]	KNN, CNN	90.3 (F1-score)	Signal
		[12]	SVM, KNN, Random Forest, etc.	80.5 (Accuracy)	Numeric
		[16]	Random Forest	58.1 (Accuracy)	Numeric
		[17]	CNN	81.8 (Accuracy)	Image
		[18]	ResNet, VGG	89.0 (Accuracy)	Image
		[19]	SVM, XGBoost, Random Forest	91.9 (Accuracy)	Numeric

다. LIME은 모델의 종류와 관계없이 사용 가능하므로 기존의 모델을 설명 가능한 모델로 변환할 수 있다. 따라서 다양한 형식의 의료 데이터 분석을 위해 LIME 기법이 사용되고 있다.

[11]의 저자들은 심혈관 질환의 치료와 예방을 위해 심전도를 해석하여 심장 박동의 종류를 분류하는 XAI의 솔루션을 제안했다. 저자들은 K-Nearest Neighbor (KNN), CNN 등의 모델을 이용하여 F1(90.3%), recall(89.5%), precision(91.9%) 및 AUC(88.0%)의 분류 성능을 측정하였다. 또한 LIME 기법을 이용해 심전도 그래프에 의사결정 근거를 표현하였다. 이를 통해 심전도 데이터 간의 시간적 의존성이 의사결정에 영향을 준다는 사실을 보였다. 또한 Support Vector Machine (SVM), Logistic Regression, KNN, Decision Tree, Random Forest 등을 활용한 Pima 당뇨병 분류 모델과 LIME 기법을 적용한 XAI 프레임워크를 제안되었다[12]. 제안하는 프레임워크를 통해 AI 모델들의 사례별 예측을 진행하였고 설명 가능한 Random Forest 모델에서 최대 80.5%의 예측 정확도를 보였다. 또한 포도당 수치가 당뇨병의 결과에 47% 정도로 가장 큰 영향을 주는 요소임을 보였다.

XAI는 크게 CAM, LRP 및 LIME 기반 기술로 나눌 수 있다. 표 1은 의료 분야에서 XAI 기술 활용 현황을 연구 현황에 따라 비교 정리한 표이다. CAM 기법은 이미지의 어떤 부분이 분류에 영향을 주었는지 설명할 수 있다. 하지만 CNN 기반 이미지 분류 모델에만 사용할 수 있기 때문에 확장성이 없다는 단점이 있다. LRP는 분해를 통해 계층별 기여도를 계산할 수 있다. CAM 기법과 달리 더욱 다양한 모델에 적용할 수 있지만 딥러닝 모델에만 적용 가능하기 때문에 딥러닝 모델에서 확장성을 가진다. LIME은 데이터의 어느 영역을 분류 근거로 사용했는지 설명한다. LIME은 위의 두 모델과 달리 모든 예측 모델에 적용할 수 있다. 따라서 모든 데이터 형식의 분류에 사용할 수 있기 때문에 모든 모델에 대하여 확장 가능성이 있다. 하지만 분류 모델의 복잡성이 미리 정의되어 있어야 한다는 단점이 있다.

Ⅲ. 결론 및 향후 연구 방향

본 논문에서는 XAI 기술을 활용한 의료 인공지능 연구들을 소개한다. 이를 통해 의료 인공지능에 XAI 기법을 적용하여 이미지, 수치 데이터 및 신호 데이터 등의 다양한 의료 데이터 형식에 따른 의사결정의 근거 및 결과 도출 과정을 설명할 수 있다. 그러나 아직 XAI의 설명 가능성에 대한 평가 기준이 명확하지 않다. 현재 설명 가능성에 대한 평가는 XAI가 제공한 정보을 사람이 개입하여 직접 평가해야 한다는 한계를 갖고 있다. 따라서 추후 XAI 모델 간의 설명 가능성 대해 비교할 수 있는 평가 기준을 설계하는 연구가 필요하다.

Ⅳ. ACKNOWLEDGEMENTS

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (No.2019-0-01601, 미래형 스마트의료 선도국으로의 도약을 위한 ICT 핵심기술 개발 및 혁신인재 양성)과 2018년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음"(2018-0-00192).

참고문헌

[1] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with

deep neural networks," Computers in biology and medicine, vol. 89, pp. 248-255, Oct 2017.
[2] K. Shankar, S. K. Lakshmanaprabu, Deepak Gupta, Andino Maseleno and Victor Hugo C. de Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification," The Journal of Supercomputing, vol. 76, pp. 1128-1143, Jul 2018.
[3] Tsigelny Igor F, "Artificial intelligence in drug combination therapy," Briefings in bioinformatics, vol. 20, pp. 1434-1448, Jul 2019
[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning deep features for discriminative localization," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921-2929, Jun, 2016
[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PloS one, vol. 10, Jul 2015.
[6] M. T. Ribeiro, S. Singh and C. Guestrin, (2016, August). ""Why should i trust you?" Explaining the predictions of any classifier," 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, Aug 2016.
[7] J. Cai, F. Xing, A. Batra, F. Liu, G. A. Walter, K. Vandenborne and L. Yang, "Texture analysis for muscular dystrophy classification in MRI with improved class activation mapping," Pattern recognition, vol. 86, pp. 368-375, Feb 2019.
[8] H. G. Nguyen, A. Pica, J. Hrbacek, D. C. Weber, F. La Rosa, A. Schalenbourg, R. Sznitman and M. B. Cuadra, "A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps," International Conference on Medical Imaging with Deep Learning, pp. 370-379, May 2019.
[9] Y. Yang, V. Tresp, M. Wunderle and P. A. Fasching, "Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks," 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 152-162, June 2018.
[10] F. Eitel, E. Soehler, J. Bellmann-Strobl, A. U. Brandt, K. Ruprecht, R. M. Giess, J. Kunchling, S. Asseyer, M. Weygrandt, J.Hynes, M. Scheel, F. Paul and K. Ritter, "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation," NeuroImage: Clinical, vol. 24, pp. 102003, 2019.
[11] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza and H. Gamboa, "Interpretable Heartbeat Classification using Local Model-Agnostic Explanations on ECGs," Computers in Biology and Medicine, vol. 133, pp. 104393, Jun 2021.
[12] P. F. Khan and K. Meehan, "Diabetes prognosis using white-box machine learning framework for interpretability of results," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1501-1506, Jan 2021.
[13] E. A. Sharara, A. Tsuji, S. Karungaru and K. Terada, "Prediction of the VDT Worker's Headache Using Convolutional Neural Network with Class Activation Mapping," IEEJ Transactions on Electrical and Electronic Engineering, vol. 15, pp. 1691-1698, Nov 2020.
[14] M. Böhle, F. Eitel, M. Weygandt and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," Frontiers in aging neuroscience, vol. 11, pp. 194, Jul 2019.
[15] F. Eitel, K. Ritter and Alzheimer's Disease Neuroimaging Initiative (ADNI), "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, pp. 3-11, Oct 2019.
[16] N. Thanh-Hai, T. B. Tran, A. C. Tran and N. Thai-Nghe, "Feature Selection Using Local Interpretable Model-Agnostic Explanations on Metagenomic Data," International Conference on Future Data and Security Engineering, pp. 340-357, Nov 2020.
[17] A. Xiang and F. Wang, "Towards interpretable skin lesion classification with deep learning models," AMIA annual symposium proceedings, vol. 2019, pp. 1246, Mar 2020.
[18] M. M. Ahsan, K. D. Gupta, M. M. Islam, S. Sen, M. Rahman and M. Shakhawat Hossain, "COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities," Machine Learning and Knowledge Extraction, vol. 2, pp. 490-504, Oct. 2020.
[19] J. Peng, K. Zou, M. Zhou, Y. Teng, X. Zhu, F. Zhang and J. Xu, "An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients," Journal of Medical Systems, vol. 45, pp. 1-9, Apr 2021.