어텐션과 어텐션 흐름 그래프를 활용한 의료 인공지능 모델의 설명가능성 연구

이유진, 채동규*
한양대학교 컴퓨터· 소프트웨어학과
{jinleey, dongkyu}@hanyang.ac.kr

A Research on Explainability of the Medical AI Model based on Attention and Attention Flow Graph

You-Jin Lee and Dong-Kyu Chae
Dept. of Computer Science, Hanyang University

요 약

의료 인공지능은 특정 진단에서 높은 정확도를 보이지만 모델의 신뢰성 문제로 인해 활발하게 쓰이지 못하고 있다. 이에 따라 인공지능 모델의 진단에 대한 원인 설명의 필요성이 대두되었고 설명가능한 의료 인공지능에 관한 연구가 활발히 진행되고 있다. 하지만 MRI 등 의료 영상 인공지능 분야에서 주로 진행되고 있으며, 이미지 형태가 아닌 전자의무기록 데이터 (Electronic Health Record, EHR) 를 기반으로 한 모델의 설명가능성 연구는 EHR 데이터 자체의 복잡성 때문에 활발하게 진행되지 않고 있다. 본 논문에서는 전자의무기록 데이터인 MIMIC-III (Medical Information Mart for Intensive Care) 를 전처리 및 그래프로 표현하고, GCT (Graph Convolutional Transformer) 모델을 학습시켰다. 학습 후, 어텐션 흐름 그래프를 시각화해서 모델의 예측에 대한 직관적인 설명을 제공한다.

1. 서론

의료 인공지능은 환자 데이터를 활용하여 질병의 진단 및 예측과 같은 맞춤형 의료 서비스를 제공하기 위해 사용될 수 있다. 그러나 환자의 생명과 직결되는 문제인 만큼, 실무에 적용하기 위해서는 의료진이 충분히 납득할 수 있을 정도의 타당한 인과 관계가 존재해야 한다. 따라서, 의료 인공지능 모델의 의사결정 과정에 대한 의료 전문가의 이해를 돕기 위해 모델에 설명력을 부여하는 연구의 필요성이 높아졌다.

의료 인공지능의 설명가능성에 관한 연구들 중 CT, MRI 와 같은 이미지 데이터로 학습된 모델은 결과와 관련된 영역을 강조 (highlighting) 하여 설명을 제공한다. 이미지 기반의 모델 설명 연구([1],[2])는 활발히 진행되고 있는 반면, 전자의무기록 (Electronic Health Record, EHR) 데이터 기반의 모델 설명에 관한 연구는 매우 부족한 실정이다. 환자 맞춤형 의료 서비스 제공을 위해서는 환자 개개인의 포괄적인 의료 기록을 활용하는 모델이 필수적인 만큼, 본 연구에서는 전자의무기록 데이터를 이용해서 모델을 학습시키고 모델의 예측 결과를 설명하는 연구를 다룬다.

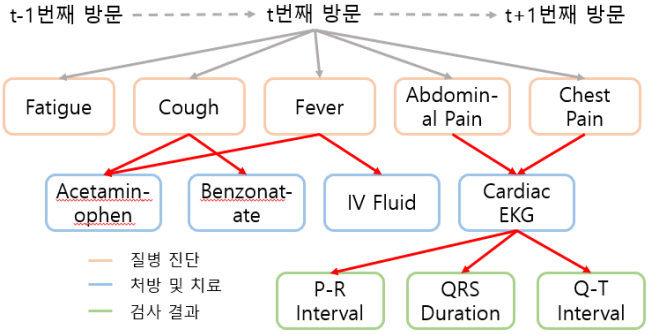
*교신저자

구체적으로, 본 논문은 GCT (Graph Convolutional Transformer) 모델에 어텐션 흐름 (Attention Flow) 를 접목하여 복잡한 의료 인공지능 예측 결과에 대한 적절한 설명을 제공하고자 한다. 데이터로는 대표적인 전자의무기록인 MIMIC-III (Medical Information Mart for Intensive Care) 를 활용하였다. 해당 데이터에 대한 전처리 및 그래프 구축, 어텐션 플로우 그래프 시각화를 통한 모델 예측의 설명 제공 방안 등을 다룬다.

2. 관련 연구

전자의무기록은 환자의 신장, 체중, 혈액형 등의 인구통계학적 정보, 의료 검사 결과, 질병 진단 정보, 처방 및 치료 정보를 포함한다. 각 정보의 상관 관계를 표현하기 위해서 그래프 구조를 활용하는 것이 효과적이다. 예를 들어 환자 A 가 병원에 방문한 시점을 t 라 하였을 때, [그림 1]과 같이 전자의무기록을 도식화 할 수 있다. 그래프는 질병 진단, 처방 및 치료, 검사 결과의 3 가지 타입(type)의 정점을 가지며, 질병 진단은 처방 및 치료로, 처방 및 치료는 검사 결과로 연결되는 단방향 간선만을 갖는다. 즉, 정점의 타입에 따라 존재하지 않는 간선이 있기 때문에 정점의 타입에 따른 계층적 구조를 가지며, 결과적으로 이형 (heterogeneous) 방향 3 계층 (3-mode) 그래프 형

태를 갖는다. 또한, 이와 같은 형태로 데이터를 구조화하면 각 그래프는 방문 시점에 따라 시간 순으로 나열되므로 시계열적 정보도 얻을 수 있다는 장점이 있다.



(그림 1) 전자의무기록 데이터의 그래프 구조.

참고문헌 [3] 에서는 전자의무기록을 그래프로 구조화하며 환자의 재입원 여부를 예측하는 GCT 모델을 제시하였다. 또한, 참고문헌 [4] 에서는 셀프 어텐션 (self-attention) 을 통과한 정보의 흐름이 많아졌을 때의 정량화 문제를 고려하여 어텐션 가중치(weights)가 주어졌을 때 입력 토큰에서 어텐션을 추정하는 방법을 제시하였다.

그러나 [3]의 모델에 사용되는 어텐션의 차원이 크기 때문에 일반적인 시각화가 어렵고 직관적이지 않아 의료 종사자에게 모델의 의사 결정 과정을 납득시키기 어렵다는 단점이 있다. 본 연구는 선행 연구의 개념을 기반으로 전자의무기록 및 MIMIC-III 와 같은 의료 데이터로 확장하고 보다 더 직관적인 설명이 가능하도록 개선하고자 한다. 이를 위해 시퀀스 데이터에 적용되었던 [4]의 개념을 그래프 데이터에 맞게 변형하고, 모델 결과를 설명하는 어텐션의 차원을 축소하여 어텐션 흐름 그래프로 시각화한다. 이러한 설명은 기존의 방법들보다 더 직관적인 인과 관계를 제공할 것으로 기대한다.

3. 제안하는 방법

본 장에서는 대표적인 전자의무기록 데이터인 MIMIC-III 를 GTC 모델 학습에 활용하기 위한 ICD-9 기반 전처리 방안, 그리고 모델 예측 결과의 직관적인 설명을 제공하는 단순화된 어텐션 시각화 방안을 다룬다.

일반적으로 관련 연구들([3],[5])에서 전자의무기록 데이터는 (예를들면, Philips eICU Collaborative Research Dataset 등) 한 환자가 병원에 입원 후 병원에 청구할 때 까지를 한 단위로 가정하고 전처리를 수행한다. 해당 방식으로 전처리된 데이터의 크기는 다양하다. 참고문헌 [3] 에서는 eICU 데이터를 GCT 모델 위에서 학습할 수 있도록 패딩을 통해 크기를 고정시킨다. 본 논문은 MIMIC-III 데이터를 동일한 방식으로 처리하기 위해 환자 식별자, 진단, 치료, 검사 결과로 크게 변수를 분리하고, 변수 별 최대 크기를 제한하여 일정한 크기의 입력을 가지도록 만들었다. 또한, 해당 데이터가 전자의무기록 시스템의 버전에 따라 다르게 입력된 값들이 있으므로 이를 동일한 의미를 가지도록 만들기 위해, 의학 온톨로지 데이터인 ICD-9 를 참고하여 값을 정규화했다. 전처리 이후 GCT 모델 학습이 수행되며, 학습과 관련된 모델 파라미터 등의 세팅은 참고문헌 [3] 과 동일하게 지정하였다. 단, eICU 데이터에 적용된 24 시간 이내 입원 조건은 MIMIC-III 에서 적용시키지 않았다.

대부분의 학습 데이터는 실제 데이터보다 큰 차원으로 투영되어 학습이 진행되기 때문에, 모델 학습 결과에 대한 어텐션 맵 (attention map) 을 시각화하면 3x101x101 의 크기가 되어 직관적인 이해가 어렵다. 따라서, 실제 데이터 길이에 따라 어텐션 행렬을 축소하고 임계값을 지정하여 필요할 경우 한번 더 필터링을 거친다. 그 다음 다시 표준화한 후 단순화된 어텐션 맵을 [그림 2]와 같이 그릴 수 있다. 또한, 전자의무기록 데이터는 진단에서 치료로 일정한 순서를 지니고 있으므로, 어텐션 흐름 [4] 을 적용함으로써 단순화된 어텐션 흐름 그래프를 [그림 3] 과 같이 시각화 하는 것이 가능하다.

4. 결과 예시

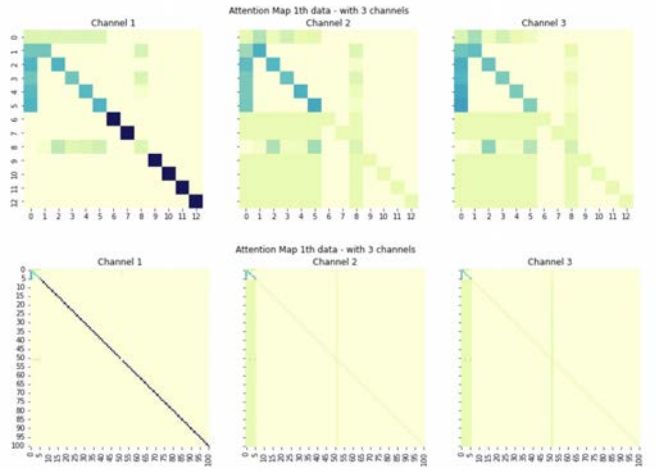
본 논문에서 사용된 데이터는 2001 년부터 2012 년까지 약 4 만명 이상의 비식별 보건의료 데이터를 포함한 중환자실 의무기록 데이터인 MIMIC-III 로 실험에는 신생아 중환자실 데이터를 제외한 후 GCT 모델 학습에 알맞게 전처리를 진행하고 학습을 수행하였다.

[그림 2] 는 GCT 의 셀프-어텐션 계산을 통해 각 블록 (block) 에서 계산된 어텐션 행렬이다. 위 3 개는 제안하는 방법을 통해 필터링된 결과이며 아래 3 개는 기존 입력 데이터 전체에 대한 어텐션 행렬이다. 제안하는 방법의 설명이 보다 더 직관적이고 내용 파악이 용이함을 확인할 수 있다.

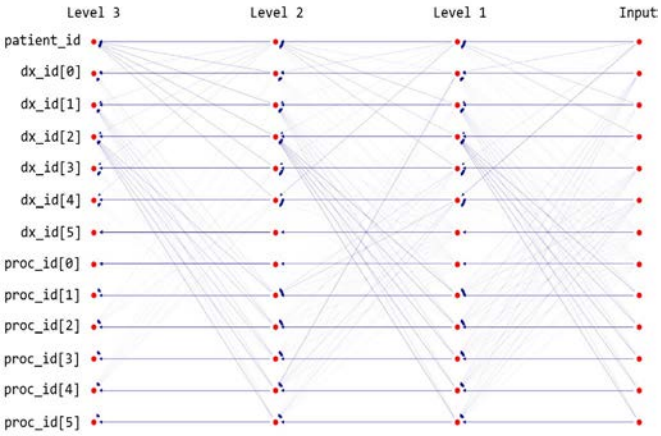
[그림 3]은 필터링 된 값들을 사용하여 어텐션 플로우 그래프로 시각화한 것으로 간선의 진하기에 따라 각 변수가 어떤 변수로 더 많이 이동했는지 정보의 흐름을 파악할 수 있다.

예를 들어, dx_id[2]의 경우 level 1~3 에서 모두 proc_id[1]~[5]의 변수와 일정하게 누적되어 연관이 있음을 확인할 수 있다. 각 치료절차와 관련된 변수는 prod_id[1] 순환기과 의사의 진료 상담, prod_id[2] 소염

진통제 오메프라졸 처방, prod_id[3] 심근허혈/심근경색으로 인한 혈관 확장제 니트로글리세린 처방, prod_id[4] 혈압강하약제 처방, prod_id[5] 베라파밀 정맥 주사로 치료를 진행했음을 의미한다. 이를 통해 dx_id[2]가 비수술적으로 치료 가능한 심혈관 질환이기 때문에 치료과정이 수술이 아닌 관련 약물 처방 및 주사 투여에 국한되어 있음을 확인할 수 있다.



(그림 2) 채널 별 어텐션 행렬 시각화.



(그림 3) 어텐션 플로우 그래프. 상세 내용은 부록 참조.

5. 결론

본 논문은 전자의무기록 데이터를 그래프 구조로 형상화하고 GCT 모델 학습, 어텐션 흐름을 접목한 시각화 등 일련의 과정으로 의료 인공지능 모델의 예측을 설명하는 연구에 대해서 다루었다. 제안하는 설명 방법이 기존의 설명에 비해 더 직관적이고 풍부한 정보를 포함하는 것을 확인하였다.

감사의 글

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 (1) 한국연구재단 바이오. 의료기술개발사업의 지원 (No. NRF-2021M3E5D2A01021156)과 (2) 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-01373,인공지능대학원지원(한양대학교))

참고문헌

[1] Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[2] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

[3] Choi, Edward, et al. "Learning the graphical structure of electronic health records with graph convolutional transformer." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 01. 2020.

[4] Abnar, Samira, and Willem Zuidema. "Quantifying attention flow in transformers." *arXiv preprint arXiv:2005.00928* (2020).

[5] Choi, Edward, et al. "Mime: Multilevel medical embedding of electronic health records for predictive healthcare." *Advances in neural information processing systems* 31 (2018).

부록

dx_id[0]	cardiovascular | vascular disorders | hypertension | malignant
dx_id[1]	cardiovascular | vascular disorders | hypertension | uncontrolled
dx_id[2]	admission diagnosis | non-operative organ systems | organ system | cardiovascular
dx_id[3]	admission diagnosis | was the patient admitted from the o.r. | no
dx_id[4]	admission diagnosis | all diagnosis | non-operative | diagnosis | cardiovascular | angina, stable
dx_id[5]	Cardiovascular | chest pain/ashd | chest pain | r/o myocardial ischemia
proc_id[0]	cardiovascular | hypertension | ace inhibitor | enalapril | iv
proc_id[1]	cardiovascular | consultations | cardiology consultation
proc_id[2]	gastrointestinal | medications | stress ulcer prophylaxis | omeprazole
proc_id[3]	cardiovascular | myocardial ischemia/infarction | nitroglycerin | intravenous
proc_id[4]	cardiovascular | hypertension | antihypertensive combination agent
proc_id[5]	cardiovascular | hypertension | calcium channel blocker | verapamil

(표 1) 어텐션 플로우 그래프 모델 변수 상세