기계학습 및 심층학습에서의 인과추론의 연구 동향

이현원 [1], 홍권 [1], 홍원석 [2], 최현수 [1,*]

[1] 서울과학기술대학교 컴퓨터공학과, [2] 강원대학교 컴퓨터공학과

{lee.hyunwon999, ghdrnjs3, *choi.hyunsoo}@seoultech.ac.kr, 4756hong@kangwon.ac.kr

A Advances of Causal Inference in Machine Learning and Deep Learning

[1]Lee Hyun-Won, [1]Hong Kwon, [2]Hong Won-Seok, and [1]Choi Hyun-Soo

[1]Seoul National University of Science and Technology, [2]Kangwon National University

요 약

본 논문은 현 인공지능 모델 학습의 약점 중 하나인 '특정 bias 에 편향적인 학습을 하여 일반화 성능이 떨어진다'는 문제점에 대해 논한다. 이에 대한 해결책 중 하나 다양한 개입이 이뤄진 입력 값에 의한 학습을 통해 보다 일반적인 인과관계를 구축 및 추론할 수 있는 방법인 인과추론의 기계학습 및 심층학습에서의 연구동향을 논한다. 더불어 인과추론의 향후 발전방향을 제시한다.

Ⅰ. 서 론

기존의 기계학습은 자신에게 부여된 작업에 대해선 성능이 좋게 나오지만, 자신이 하는 일에서 약간만 벗어난 주제를 질문하면 성능이 크게 하락하는 모습을 보이며, 이는 현재 인공지능이 아직 인간을 넘지 못한 분야 중 하나로 여겨진다.[1] 이는 모델의 일반화 성능에 대한 문제와 직결되며, 가령 합성곱 신경망 분류 문제에서 코끼리 이미지의 픽셀 값을 조금만 수정하여도 잘 동작하는 모델이 코알라로 분류하는 등의 예시가 있다. 이렇게 강인하지 못한 모델은 사용자에게 요구되는 작업별로 다른 모델을 만들어야 하는 등, 재사용성이 떨어지는 점으로 인해 기존의 학습된 지식을 활용하지 못하는 문제가 있다. 따라서, 장기적인 관점에서 사람처럼 기존에 학습한 지식을 일반화 시켜 저장하고 이를 다른 작업 시에 활용할 수 있게끔 하는 능력이 필요하다고 할 수 있는데, 이 작업을 위한 '명시적 일반화'에 해당하는 개념이 인과추론이다. 이것에 관한 주제로 작성된 다른 논문들을 통해, 기계학습 분야에서 인과추론의 발전사항과 향후 적용 가능한 도메인에 대해 논하도록 한다.

Ⅱ. 본 론

A. 인과추론의 개념

기계학습은 통계적인 방식으로 입력 데이터를 학습하는데, 정해진 작업에 맞는 입력 데이터를 사람이 직접 전처리하고, 해당 작업에 맞는 정답 값을 주거나 그에 준하는 목표를 제시하며, 모델이 제시한 값과 실제 정답 값과의 차이를 미분하여 모델이 실제 정답 값을 출력하도록 학습하는 방식이다.[1] 단, 이렇게 작업하는 경우 입력 데이터는 일관적인 형태를 유지해야 하는데, 입력 데이터에 대한 개입으로 인해 일관성에 변형이 일어나는 경우 모델이 학습에 사용한 요소인 '기울기' 값이 초기값(통제된 외부변수)에 의존하는 만큼 성능 측면에서 하락을 일으키기 때문이다. 이렇게 목적에 따라 입력 데이터의 분포를 적절히 통제한다는 개념을 i.i.d.(independent and identically distribution; 독립항등분포) 가정이라 한다. 가령 황새와 출산율에 대한 예측 모델을 세우는 경우, 해당 요소들에 편향적인 데이터를 입력으로 준다면 그 모델은 그 실험 조건 내에서만 정확한 값을 내도록 훈련될 것이다.[1] 따라서 실험 조건을 벗어나는 상황에서도 기존에 학습한 지식을 활용할 필요가 있는데, 그 방법 중 하나로 인과추론이 있다. 그림 1 에서 볼 수 있듯이, 학습 시에 입력 데이터로부터 다양한 개입(입력 값의 변화)을 통해 보다 일반적인 관점에서의 인과구조를 학습하도록 하며, 나아가 해당 정보들을 바탕으로 입력 데이터에 대한 편향되지 않은 지식을 구축할 수 있도록 돕는다.

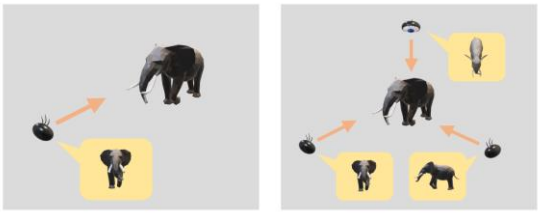

그림 1. bias 와 개입에 대한 추상적 표현. (a) 기존에 수행되는 딥러닝 방법론의 bias, (b) 개입을 통해 bias 에 치중하지 않고 다양한 관점을 통해 데이터를 바라보는 모습

B. 인과추론의 발전 동향

이에 대한 발전 과정은 그림 2 와 같은데, 먼저 2000 년에 J. Pearl 의 저서[2]에서 인과 사다리, SCM(Structural Causal Models; 구조적 인과모델) 등의 기초적인 인과추론 개념들이 등장했다. SCM 은 관측 가능한 변수 집합을 고려하기 위한 인과 모델이며, 수식으로 정의되어 있다.[1]

$$X_i := f_i(PA_i, U_i), \quad (i = 1, \ldots, n) \qquad (1)$$

여기에서 PA_i는 그래프이며, 인과 요소 X_i의 부모값에 따른 결정론적 함수 f_i를 사용하고, 설명할 수 없는 랜덤 변수를 U_i로 칭한다. 인과추론의 목표는 인과관계가 복잡하게 얽힌 입력값으로부터 일반적이고 독립적인 인과 요소 X 를 추론해내는 것이다. 이러한 일련의 인수분해 과정의 수식은 다음과 같다[1] :

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid PA_i) \qquad (2)$$

그러나 해당 이론을 컴퓨터 알고리즘으로 구현하기엔 당시 하드웨어 성능과 효과적인 프레임워크가 없었기에 유의미한 성과를 나타낸 연구가 부족했다. 이후 컴퓨터의 성능이 향상되면서 기계학습에 대한 연구가 가속화되었고, 2016 년에 Johansson 의 논문[3]을 통해 인공지능 모델의 반사실적 질문에 대답하는 작업 등의 인과적 추론을 위한 표현 학습에 대해 다루며, 딥러닝에서의 인과추론의 초석을 놓은 것으로 인식되었다. 이후 SCM 원칙을 만족하는 인과추론 모델을 만들기 위한 다양한 시도가 있었는데, 한 예시로 Propensity-Dropout 을 사용하여 selective bias 를 없애는 시도[4]가 있다. 기존 딥러닝에서 과적합을 막기 위해 사용하던 dropout 에서, 특정 노드의 배제 확률을 추가적인 Propensity Network 를 추가 및 학습시킴으로써 표본의 편향을 줄인다는 개념이다.

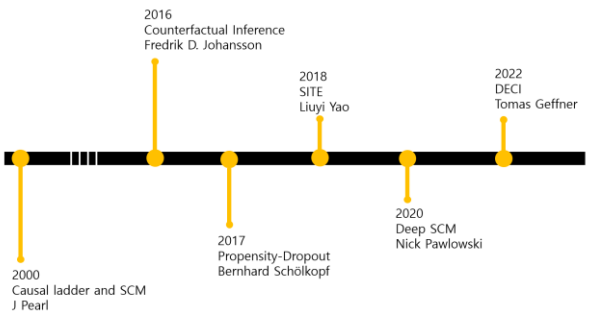

그림 2. 인과추론 타임라인

인과추론은 의료 분야에서도 면밀히 다루는 대상인데, 이 또한 대조군과 치료군 분포의 균형을 맞추기 어려운 selective bias 가 존재하기 때문이다. 이에 따라, 인과 그래프 모델을 이용하여 국지적 유사성과 데이터 분포의 균형을 동시에 유지하는 모델을 제안하였다.[5] 이러한 연구 성과들에 힘입어, 2020 년에는 심층 SCM 훈련을 위한 프레임워크가 제안되었는데[6], Pearl 이 제안했던 인과 사다리의 세 단계인 association, intervention, counterfactuals 모두를 수행 가능하게끔 만들어졌고, 2022 년에는 개별적으로 연구되던 인과관계 발견과 인과추론 분야를 동시에 수행할 수 있는 알고리즘인 DECI[7]가 만들어졌다. 특히 기존의 의학 분야에서 사용되던 평균치료효과(Average Treatment Effect, ATE)에서 특정 하위 집단 혹은 조건에 따라 달라지는 치료효과인 CATE(Conditional ATE)의 계산에 인과 그래프를 접목시키는 시도로, 이러한 의학적인 관점에서의 인과추론이 그 필요성에 의해 활발히 연구되고 있다.

Ⅲ. 결 론

본 논문에서는 기계학습의 i.i.d. 가정에 의한 약점과 그것을 해소하기 위한 방안 중 하나인 인과추론에 대해 소개하였으며, 복잡하게 얽혀 있는 인과관계로 구성된 입력 데이터로부터 독립적인 인과를 추론해낼 수 있는 모델 구성의 의의와 그 장점을 소개하고 있다. 이 방향성에 기반하여, 기계학습에 인과추론 모델을 사용하는 경우에 대한 몇 가지 논의점을 제시할 수 있다.

분리된 표현 학습: 복잡하게 얽힌 입력 데이터로부터 인과관계를 추출할 때, 최대한 각각을 독립적으로 분리해야 할 것이다. 그러나 그것을 어느 수준까지 분리할지, 혹은 어느 상황에 어느 수준의 인과관계를 살펴볼지 등을 task 에 따라 나누는 것은 일반화 성능을 떨어뜨릴 수 있으므로, 이를 만족하는 인과관계 생성 및 저장 방식이 필요하다.

변형 가능한 메커니즘 학습: 인과관계 발견 및 학습이 특정 task 에만 국한되는 식으로 경직된다면, 이는 인과추론의 관점에선 bias 가 들어갔다는 점으로 인해 다른 task 에 적용하지 못하는 문제가 생길 것이다. 따라서 특정 task 에 맞도록 모델을 학습시키더라도, add-on 형식으로 장착 가능한 인과추론 모델[1] (Schölkopf 2021, 7)은 입력값의 bias 등의 변화하는 배경 조건에서도 불변성을 유지할 수 있도록 별개의 학습이 되어야 할 것이다.

강화학습에서의 인과추론: 적절한 인과관계를 형성한 경우, 모델 스스로 '과거의 다른 선택이 현재 어떤 결과를 불러왔는지'와 같은 counterfactual 상황을 상정함으로써 더 나은 선택을 하게끔 학습할 수 있을 것이다. 이것은 강화학습 모델 (e.g., Deep Q-Network)의 experience replay 와 그 궤를 같이할 수 있으므로 서로 상보관계에 있다고 할 수 있을 것이다. 특히 인과추론에 대한 최근 연구가 의료 도메인에서 주로 진행되어온 만큼, 다른 분야로의 확장에 인과추론이 강화학습의 형태로 융합되는 것을 기대할 수 있다.

ACKNOWLEDGMENT

본 논문은 한국과학기술연구원 기본사업 (2E32260 and 2E32264) 및 한국연구재단 이공계기초연구사업 (NRF-2022R1F1A1076454)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] Schölkopf, Bernhard, et al. "Toward causal representation learning." Proceedings of the IEEE 109.5 (2021): 612-634.

[2] Pearl, Judea. Causality. Cambridge university press, 2009.

[3] Johansson, Fredrik, et al. "Learning representations for counterfactual inference." International conference on machine learning. PMLR, 2016.

[4] Alaa, Ahmed M., Michael Weisz, and Mihaela Van Der Schaar. "Deep counterfactual networks with propensity-dropout." arXiv preprint arXiv:1706.05966 (2017).

[5] Yao, Liuyi, et al. "Representation learning for treatment effect estimation from observational data." Advances in neural information processing systems 31 (2018).

[6] Pawlowski, Nick, et al. "Deep structural causal models for tractable counterfactual inference." Advances in Neural Information Processing Systems 33 (2020): 857-869.

[7] Geffner, Tomas, et al. "Deep end-to-end causal inference." arXiv preprint arXiv:2202.02195 (2022).