합성 데이터와 통제된 실험 환경

이동재

20240905

Abstract

합성 데이터는 단순히 데이터를 증강하는 역할을 넘어 인공지능의 성능 실험을 위한 통제된 데이터를 구축하는데 사용될 수 있다. 기존 인공지능 학습 데이터는 무분별하게 수집된 데이터로, 어떤 정보가 포함되어 있는지 알 수 없다. 반면 합성 데이터는 생성자가 학습 데이터 속 정보의 총량, 종류, 분포를 정확히 알고 통제할 수 있다. 본 글에서는 합성 데이터가 인공지능의 분석에 어떻게 활용될 수 있는지 소개하고자 한다.

합성 데이터는 인간이 생성한 데이터가 아니라, 인공적으로 만들어진 모든 데이터를 지칭한다. 합성 데이터는 언어 모델, 확률 문법, 비결정적 상태 기계 등 여러 종류의 모델로부터 무작위로 추출하여 생성한다. 이러한 기법은 복잡한 추론 문제를 만들어내는데 유용하게 사용되며, 수학, 코딩 등 논리적 추론 능력을 향상하는데 도움이 된다.

합성 데이터는 인공지능을 분석하고, 성능을 평가하는데도 사용될 수 있다. 기존 인공지능의 성능 평가는 평가 벤치마크에 초점이 맞춰져 있으며, 학습 데이터에 대한 통제는 미미하다. 이 때문에 평가 벤치마크에 있는 데이터를 이미 학습 데이터에서 학습했을 가능성이 항상 존재한다. 합성 데이터는 이러한 한계를 극복하는 방법으로, 생성자가 데이터를 직접 생성하기 때문에 어떤 정보가 담겨있는지 정확히 알 수 있다.

이러한 아이디어를 기반으로 '언어 모델이 진정으로 추론 능력을 지니고 있는가?'에 답하는 실험을 소개한다. 이 실험에서는 무작위로 비순환 유향 그래프를 생성하고, 이를 기반으로 복잡한 추론 문제를 만들어내 언어 모델의 성능을 평가한다. 그 결과, 언어 모델은 데이터를 통해 추론 자체를 학습할 수 있음을 발견하였다. 학습 데이터부터 통제된 실험 환경이었기 때문에 분포 밖 데이터에 대한 성능 평가도 가능하였고, 이것이 추론 능력의 결정적 증거가 되었다. 기존 방식으로는 무엇이 분포 밖 데이터인지 알 수 없었기 때문에 완벽히 통제된 실험은 불가능했을 것이다.

통제된 실험은 과학에서 가장 중요한 요소지만, 데이터를 수집하기 어렵다는 이유로 학습 데이터에 대한 통제는 경시되고 있다. 합성 데이터를 통해 학습 데이터를 구축하고, 이를 기반으로 신경망을 분석한다면 보다 과학적인 방법으로 인공지능을 연구할 수 있을 것이다.