< 고품질 데이터 중심의 AI 연구/개발 필요성 >

정보컴퓨터공학과 202055565 여지수

『 인공지능의 한계 3 - AI 응용 상용화 85% 실패, 그리고 Data-Centric AI로의 이동 』

필자는 머신러닝 개발자를 희망하는 컴퓨터공학과 학부 3학년 재학생이다. 대학교에서 머신러닝 분야와 관련된 수업을 들으면서 느낀 것은 방대한 데이터가 머신러닝에서 제일 중요한 점이라는 것이다. SAMSUNG SDS의 인사이트에서 AI 분야 전문가들의 글을 읽던 중 '데이터 중심의 모델 연구'를 펼쳐야한다는 한 인사이트 내용에 감명을 받아 이를 바탕으로 글을 작성하게 되었다.

AI는 우리 생활의 많은 부분들에 자리를 잡고 있고 꾸준히 황금기를 이어갈 것 같지만, 사실 AI는 여러 한계들을 가지고 있다. 그 중 "데이터" 측면에서 몇가지를 이야기 해보겠다. 기계를 학습시키기 위해서는 방대한 양의 데이터가 필요하다. 그러나, 자연의 셀 수 없는 데이터들을 수집하는 것은 쉽지 않은 일이다. 또, 데이터들의 분포는 시간에 따라 변하기 때문에 현재로써의 학습 이론에는 한계가 존재한다. 이러한 한계들을 해결하려면 데이터들이 방대하지 않더라도 신경망을 정확히 학습하는 이론이 필요하다. "오염된 데이터셋을 보정하여 Clean Labels로 만드는 기술", "품질이 좋은 새로운 데이터를 새로 자동 생성하는 기술" 을 연구하는 것이 앞으로 AI를 연구하는 사람들에 있어서의 숙제가 될 것이다.

인사이트 글을 통해 필자의 생각은 바뀌었다. '방대한 데이터'의 수집 보다는 '고품질 데이터'의 생성이 더 중요하다. 이는 AI 기술의 연구 뿐만 아니라 모델을 개발하는 때에도 중요하게 작용하는 부분이라 생각한다. 모델에 학습시키고 훈련시킬 데이터들을 수집할 때에 수집 후 활용 가능한 데이터를 '고품질화'하는 것에 많은 힘을 쏟아야한다. 무조건적인 방대한 데이터를 수집하는 데에 초점을 두기 보다는 학습시키기 좋은 데이터들로 정제를 하는 것에 더 초점을 두고 개발을 해야한다는 것이다. 그렇게 된다면 현 AI의 한계를 극복하고 정확도가 높은 모델을 개발 할 수 있을 것이다.

---

1 https://www.samsungsds.com/kr/insights/data_centric_ai.html