

헤드라인

AI, 년 어디까지 왔니

기사링크

<https://n.news.naver.com/article/newspaper/015/0004496550?date=20210209>

[✓ 4100 억개 자연어 처리 ✓ 텍스트& 이미지로 새 콘텐츠 생성
✓ 올해 주목할 인공지능 9 대 기술]

“GPT-3(초대규모 자연어 처리 인공지능 모델)는 자연스러운 문장을 구사한다. 하지만 인간의 뉴런에 비하면 여전히 매우 작은 규모다. 데이터, 신경망에서 더 큰 규모의 증가가 나온다면 인간의 이성적 사고와 비슷한 AI 구현이 가능해질 것이다.”

구글의 AI 를 총괄하는 제프리 힌튼 캐나다 토론토대 교수가 지난해 11 월 MIT 테크놀로지리뷰를 통해서 한 말이다. GPT-3 는 인간의 뉴런과 비교하면 여전히 빈 곳이 많지만 잠재력만큼은 파괴적이다. 학습시키는 데 투입된 데이터(단어) 수만 놓고 봐도 상징성이 크다. 크롤링(긋어오기)을 통해 모은 단어만 4100 억 개에 달한다. 그만큼 장족의 발전을 했다는 얘기다.

최근 한국지능정보사회진흥원(NIA)은 2021 년 주목해야 할 인공지능 기술로 ‘인류 역사상 최고의 언어 인공지능’으로 꼽히는 GPT-3 를 포함해 트랜스포머(transformer·병렬 처리), 자기지도학습(self supervised learning), 연합학습(federated learning), 에지 AI(edge AI), 전이학습(transfer learning), 생성적 AI(generative AI), Auto ML, 시스템 2 AI(system2 AI) 등을 꼽았다. 이들은 모두 한 방향을 가리키고 있다. AI 의 진화 속도에 급가속이 붙었다는 평가다.

본문



■ 초대규모 모델의 부상

2020 년 6 월 등장한 GPT-3 는 자연어 처리 모델의 혁명적인 스케일업을 가져왔다. GPT-3 이후 세계 테크기업들은 초대규모 모델의 우수한 성능을 인식하기 시작했다. 네이버는 지난해 10 월 GPT-3 를 능가할 초대형 한국어-일본어 모델을 개발하겠다고 선언했고, GPT-4 등 후속 모델 개발도 이어질 전망이다.

이런 초대규모 인공지능 모델의 등장 뒤에는 트랜스포머 병렬 처리 기술과 자기지도학습이 있다. 트랜스포머를 통한 병렬연결 덕분에 짧은 시간 초대용량 데이터셋의 입력이 가능해졌다. 2020 년 하반기부터는 자연어 처리 분야 외에 컴퓨터 영상 인식 분야에서도 병렬 처리 기술이 적용되면서 새로운 성과가 예상되고 있다.

자기지도학습은 입력 값의 일정 부분만으로 입력 값의 다른 부분을 예측하며 자체적으로 학습해 나가는 기술이다. 기존에는 지도학습의 원천이 되는 데이터를 수작업으로 라벨링(이름 붙이기)해야 했다. 노동 집약적인 데이터 라벨링 작업에서 해방되면서 초대용량 원시 데이터를 인공지능 모델이 학습할 수 있게 됐다.

■ 응용서비스 위한 기반 기술 주목

원천기술 측면에서 인공지능 모델이 초대규모화하고 있다면, 일상생활과 밀접한 응용 서비스 분야에서는 데이터와 모델을 경량화하는 기술이 발전하고 있다.

대표적인 분야가 연합학습과 에지 AI 다. 연합학습은 데이터를 중앙에 모아서 학습시키는 기존의 통합학습과 달리 각 기관과 기기에서 학습한 결과값만 중앙으로 취합하는 방식이다. 환자 개인정보 보호가 중요한 의료 분야에서 대규모 정밀 의학을 가능하게 하는 요소로 적극 도입되고 있다. 엔비디아는 환자의 데이터를 공유하지 않고 각 병원에서 학습하고 다른 병원과 협력이 가능한 '엔비디아 클라라 연합학습'을 최근 공개했다. 희귀질환 모델 구축, 신약 개발을 통한 출시 시간 단축 등의 혁신이 예상되고 있다.

에지 AI는 적은 메모리와 계산 능력으로 구동되는 스마트폰, 차량, 드론 등 에지 장치에서 AI를 구현하기 위한 기술이다. 성능 저하 없이 모델의 크기를 압축하는 것이 핵심이다. 최근 인공지능 모델의 정확도를 높이기 위해 데이터와 모델의 규모가 증가하고 있지만 산업 현장에서 즉시 활용은 어렵다. 이에 고성능 모델이 등장해 성능 기록을 경신하면, 이와 같은 수준의 성능을 기록하는 경량화 모델 개발이 동시에 진행되고 있다.

전이학습 역시 대표적인 경량화 기술이다. 예를 들어 구글에서 고성능 컴퓨팅 성능을 바탕으로 최고 성능의 인공지능 원천 모델을 공개하면 스타트업, 중소기업들이 해당 모델의 일부를 가져와(전이) 자신들의 모델을 학습시키는 기술이다. 앞서 언급된 GPT-3 역시 API(응용프로그램밍)를 통해 다양한 자연어 처리 인공지능 모델의 플랫폼으로 진화하고 있다.

■ AI 창조 & 추론...우수인재 의존도 높아져

생성적 AI와 Auto ML, 시스템 2 AI와 같은 차세대 AI 개발을 위해 우수한 인재에 대한 의존도가 더 높아졌다는 것이 NIA의 판단이다.

생성적 AI는 인공지능이 텍스트와 이미지 등 기존 콘텐츠를 사용해 자체적으로 새로운 콘텐츠를 생성하는 것이다. 핵심 원리는 2014년 이언 굿펠로가 소개한 '생성적 적대 신경망(GAN)'이다. 이는 실제 데이터와 비슷하게 새로운 것을 만들어내는 '생성자'와 만들어진 것을 평가하는 '판별자'가 끊임없이 서로 대립하고 경쟁하며 완벽성을 추구해나가는 방식이다. 진짜와 가짜의 싸움을 통해 진짜 같은 가짜를 복제하는 딥페이크가 이런 원리로 생산된다.

Auto ML은 머신러닝 개발 과정에서 소모적이고 반복되는 작업을 자동화하는 프로세스다. 데이터 수집부터 전처리, 모델링, 훈련, 평가 등 개발자가 하는 일을 알고리즘 최적화를 통해 이뤄낸다. 'AI를 만들어내는 AI'로도 불린다.

시스템 2 AI는 인공지능이 학습되지 않은 상황에서도 맥락을 이해하는 AI를 설명하는 개념이다. 신경망 언어모델을 통한 딥러닝 자연어 처리의 핵심 기술을 창안한 요슈아 벤지오 몬트리올대 교수가 인공지능 모델에서 추론의 중요성을 강조하면서 주목받고 있다.

NIA는 이처럼 창조와 추론이 가능한 차세대 AI를 만들어 내기 위해 "알고리즘 분야 경쟁력을 확보해야 한다"고 지적한다.

[인공지능 기술]

(1) 대규모 모델

i. 트랜스포머 (transformer·병렬 처리)

- 병렬연결 덕분에 짧은 시간 초대용량 데이터세트의 입력이 가능
- 2020년 하반기부터 컴퓨터 영상 인식 분야에서도 병렬 처리 기술이 적용

ii. 자기지도학습(self supervised learning)

정리

- 입력 값의 일정 부분만으로 입력 값의 다른 부분을 예측하며 자체적으로 학습하는 기술
- 기존에는 지도학습의 원천이 되는 데이터를 수작업으로 라벨링 했음
- 데이터 라벨링 작업에서 해방되면서 초대용량 원시 데이터를 인공지능 모델이 학습할 수 있음

(2) 응용 서비스 위한 기술

: 데이터와 모델을 경량화 하는 기술이 발전

i. 연합학습 (federated learning)

- 데이터를 중앙에 모아서 학습시키는 기존의 통합학습과 달리 각 기관과 기기에서 학습한 결과값만 중앙으로 취합하는 방식
- 환자 개인정보 보호가 중요한 의료 분야에서 대규모 정밀 의학을 가능하게 하는 요소로 적극 도입
- Ex) 엔비디아
 - ⇒ 환자의 데이터를 공유하지 않고 각 병원에서 학습하고 다른 병원과 협력 가능
 - ⇒ 희귀질환 모델 구축, 신약 개발을 통한 출시 시간 단축 등의 혁신 예상

ii. 에지 AI (edge AI)

- 적은 메모리와 계산 능력으로 구동되는 스마트폰, 차량, 드론 등 에지 장치에서 AI 를 구현하기 위한 기술
- 성능 저하 없이 모델의 크기를 압축하는 것이 핵심
- 현재 데이터와 모델의 규모가 증가하고 있지만 산업 현장에서 즉시 활용은 어려움

iii. 전이학습 (transfer learning)

- 대표적인 경량화 기술
- 구글에서 고성능 컴퓨팅 성능을 바탕으로 최고 성능의 인공지능 원천 모델을 공개하면 스타트업, 중소기업들이 해당 모델의 일부를 가져와(전이) 자신들의 모델을 학습시키는 기술

(3) 차세대 AI

i. GPT - 3 (초대규모 자연어 처리 인공지능 모델)

- 인류 역사상 최고의 언어 인공지능
- 크롤링(긋어오기)을 통해 모은 단어만 4100 억 개
- 데이터, 신경망에서 더 큰 규모의 증가가 나온다면 인간의 이성적 사고와 비슷한 AI 구현이 가능

ii. 생성적 AI(generative AI)

- 인공지능이 텍스트와 이미지 등 기존 콘텐츠를 사용해 자체적으로 새로운 콘텐츠를 생성
- 핵심 원리: '생성적 적대 신경망(GAN)'
 - ⇒ 2014 년 이언 굿펠로가 소개
 - ⇒ 실제 데이터와 비슷하게 새로운 것을 만들어내는 '생성자' 와 만들어진 것을 평가하는 '판별자' 가 끊임없이 서로 대립하고 경쟁하며 완벽성을 추구해 나가는 방식

iii. Auto ML

- 머신러닝 개발 과정에서 소모적이고 반복되는 작업을 자동화하는 프로세스
- 데이터 수집부터 전처리, 모델링, 훈련, 평가 등 개발자가 하는 일을 알고리즘 최적화를 통해 이뤄냄

iv. 시스템 2 AI(system2 AI)

- 인공지능이 학습되지 않은 상황에서도 맥락을 이해하는 AI