The promise and pitfall of synthetic data in training AI system

합성 데이터의 시대가 오고 있다

신뢰도 높은 AI 시스템을 구축하려면 양질의 데이터가 필요하지만 AI 모델의 훈련에 필요한 데이터를 구하기는 쉽지 않다. 원본 데이터의 통계적 변수 분포와 상관관계 등을 모방한 합성 데이터(재현 데이터)는 고질적인 데이터 병목현상을 해소할 수 있다.

최은창

2022년 10월 24일

데이터 부족의 문제

AI 개발자들은 몇 가지 골치 아픈 이슈들에 직면해 있다. 우선, AI 개발 과정에서 기업들은 데이터를 절실히 필요로 한다. 올바른 데이터 구하기는 강력한 AI를 구축하는 데 가장 중요하면서도 가장 어려운 부분이다. 예컨대, 의료용 AI 개발자가 고품질의 병변 데이터를 구하기는 어렵다. 이러한 데이터 공급의 한계는 AI의 발전 속도를 느려지게 만드는 요인이다.

둘째, 데이터 품질이 낮거나 데이터 세트에서 개인 정보가 노출되는 문제가 종종 발생한다. 데이터 품질의 문제는 AI 모델의 판단이 편향되거나 공정(fair)하지 않을 수 있다는 불신으로 번지게 된다. “쓰레기 데이터를 넣는다면 쓰레기가 나온다(garbage in, garbage out)”는 격언은 실무에서 여전히 유효하다. 질 낮은 데이터는 AI 모델의 연산을 거친 결과값을 신뢰할 수 없게 만든다.

셋째, AI 모델에 데이터를 공급할 때 데이터의 원본에서 개인 정보를 제거하고 사회적 불평등 논란을 미리 방지해야 한다. 정확한 예측 결과를 제공하려면 데이터 세트가 편향되지 않아야 하고 강화된 개인 정보보호 규정까지 준수해야 한다. 데이터를 구하기가 어려워지자 많은 기업들은 합성 데이터(synthetic data)에 주목하기 시작했다. 합성 데이터를 사용하면 훨씬 빠르고 적은 비용으로 AI 모델의 훈련 데이터를 확보할 수 있다. 오늘날 세계에서 가장 귀중한 자원은 데이터인데, 이 데이터를 무한한 양으로 저렴하고 빠르게 생산할 수 있는 방법이 있다면 기업의 입장에서는 관심을 가질 수밖에 없다.

합성 데이터는 어떻게 생성되나?

MIT 테크놀로지 리뷰는 2022년 10대 미래 기술 중의 하나로 합성 데이터를 선정했다. 합성 데이터는 ‘재현 데이터’로도 불리는데 실제 데이터 세트에 존재하는 통계 패턴을 모방한 데이터(simulated data)를 의미한다. 합성 데이터에 대하여 유럽 데이터 보호 감독기구(EDPS)는 “원래 데이터 소스를 가져와서 유사한 통계 속성을 가진 새로운 인공 데이터를 생성”하는 것이라고 정의한다. 간단히 말해, 합성 데이터는 데이터의 통계적 특성을 모방하여 만들어진 인공적으로 만들어진 가짜 데이터이다.

그렇다면 합성 데이터는 어떻게 만들어지는 것일까? 현실 세계의 사건들을 실제로 수집하지 않고 컴퓨터 시뮬레이션이나 알고리즘이 생성한다. 즉, 소량으로 수집된 원본 데이터 세트를 샘플로 삼아서 그 통계적 특성을 모방하여 인위적으로 만들어진다.

신경망을 이용한 합성 데이터의 생성은 생성적 적대 신경망(GAN:Generative Adversarial

Networks)에서 힌트를 얻았다. 캐나다 몬트리올대에서 요슈아 벤지오의 지도 아래 박사과정을 마친 이언 굿펠로우(Ian Goodfellow)는 동료들과 이미지를 정밀하게 그리는 기계를 개발하려는 아이디어에 대해 이야기를 나누었다. 그는 적대적 양방향의 신경망들이 서로 경쟁하면서 가짜 이미지를 사실적으로 정확하게 생성하는 방법을 생각해 냈다. 합성 데이터를 인위적으로 생성하는 방법은 생성적 적대 신경망의 작동방식으로부터 자연스럽게 얻어진 것이다.

합성 데이터의 장점은 무엇일까?
AI 모델을 훈련시킬 때는 정확히 레이블(label)된 풍부한 데이터 세트가 필요하다. 더욱 다양한 데이터로 훈련한다면 더 높은 정확도를 달성할 수 있지만 수백만 개의 대규모 데이터를 수집하고 레이블을 지정하는 작업에는 막대한 시간과 비용이 필요하다. 게다가, 실제 데이터(real-world data)는 AI 모델의 훈련에 적합하지 않은 경우도 많다. 직접 관찰하여 얻은 데이터보다 합성 데이터가 가치 있다니 언뜻 보기에 모순적으로 들릴 것이다.

실제 데이터는 물론 좋은 통찰력을 제공하지만, 실제 데이터는 우연에 좌우되는 경우가 많고 현실 세계에서 가능한 모든 조건이나 사건의 순열을 포함하지 않는다. 게다가 실제 데이터는 개인정보 보호 규정으로 인해 데이터 전처리 과정(preprocessing)에 비용이 많이 들고 엉망인 상태(messy)이거나 오염된 경우가 많다.

실제 데이터에는 부정확한 요소들과 편향(bias)까지도 포함되어 있기 때문에 데이터 정제(data cleansing) 과정을 거치지 않으면 신경망에 오히려 악영향을 미칠 수가 있다. 즉, 실제 데이터를 수집한 이후에는 데이터 전처리 과정을 거쳐서 개인 정보를 제거하고, 오류를 걸러내고 서로 다른 데이터 형식들도 통일해야만 한다. 이 과정은 번거롭고 비용을 증가시킨다.

저명한 벤처 투자가 롭 토우(Rob Toews)는 이 문제를 정확히 지적했다. AI 모델을 구축하는 많은 기업들이 이미지 레이블링 작업에 해마다 수천만 달러가 넘는 돈을 쏟아붓고 있지만 그것이 실제로100% 정확하리라는 보장이 없다고 말한다. 반면, 합성 데이터를 사용하여 AI모델을 훈련시킨다면 더욱 더 균일한 데이터 형식과 레이블을 유지할 수 있으므로 실제 데이터의 효과적인 보완책이나 대안으로 여겨지고 있다.

훈련용 데이터의 공급
합성 데이터는 레이블이 지정된 훈련 데이터를 거의 무제한으로 생성하여 심층 신경망(DNN, Deep Neural Network)에 공급할 수 있다. 또한 합성 데이터를 사용한 심층 신경망 훈련은 실제 데이터로 훈련시키는 경우에 비해 더 적은 비용이 들어간다. 합성 데이터의 확장성도 매력적인 요소이다. 머신러닝 모델을 훈련하고 테스트하는 데 필요한 데이터를 확보하는 일은 까다로운 과정이 요구되지만 합성 데이터는 생성과 사용이 더 간단하다.

예컨대, 간 병변 데이터 이미지만가 적은 분량만 있다면 질병 진단용 AI 모델을 훈련시켜도 예측의 정확도가 떨어지는 문제가 있었다. 그런데 의료용 합성 데이터를 AI 모델의 훈련에 추가적으로 사용하자 진단용 신경망의 분류 성능이 확연히 개선되었다. GAN을 사용하여 실제 데이터와 같은 가상의 의료용 영상 데이터를 생성하게 되자 환자들이 질병의 진단을 위해 방사선 검사를 받을 필요성도 크게 줄어들었다.

의료용 데이터는 민감하므로 신원이 노출되어서는 안되지만 합성 데이터는 개인정보 보호와 데이터 유용한 활용 사이에서 균형을 찾을 수 있는 방법이다. 이러한 맥락에서, 옥스포드대 바이오의료 공학과의 엘리슨 노블(Alison Nobel)은 합성 데이터가 민감한 의료 정보의 공정이용(fair use)을 확대하는 역할을 할 것이라고 전망한다.

합성 데이터 스타트업들
최근에는 개인정보 유출에 대한 우려가 커졌고 데이터 기반 솔루션 수요도 늘어났다. 이런 흐름은 합성 데이터의 활용도를 높이는 배경이 된다. 합성 데이터는 원래 데이터 세트의 통계적 특성 또는 분포를 유지하면서도 익명성을 유지시켜 준다.

합성 데이터를 다양한 용도를 위해서 제공하는 스타트업들은 2021년부터 주목을 끌고 있다. 샌프란시스코에 위치한 '신세시스 AI'(Synthesis AI)는 데이터 다양성(data diversity)에 초점을 두고 주문형 합성 데이터를 기업들에게 제공한다. 수만 명이 넘는 특이한 사람들, 개체들, 환경을 결합하여 거의 무한에 가까운 데이터 가변성을 생성한다. 컴퓨터 비전용 합성 데이터는 증강 현실(AR)과 가상 현실(VR)에서 디지털 아바타의 자세와 동작 인식 등에도 사용될 전망이다.

이스라엘 기업 데이터젠(Datagen)은 합성 데이터가 점차 컴퓨터 비전 시스템의 핵심을 담당할 것으로 내다보고 있다. 이 업체는 AI 개발자들에게 합성 데이터 생성 플랫폼을 통해 컴퓨터 비전 모델의 훈련을 돕는다. 동작 기반 시퀀스의 합성 데이터 세트는 객체 감지 및 식별에 사용된다. 딥러닝 기반의 컴퓨터 비 모델이 피부색에 따라서 안면 인식률에 큰 차이를 보인다면 인종차별 논란이 불거지게 된다. 그런데 다양한 인종의 얼굴들을 포함한 합성 데이터로 훈련시킨다면 피부색에 따른 인식률 차이는 크게 줄어들게 된다

한편, 엔터테인먼트 콘텐츠의 제작에도 합성 데이터가 사용된다. 3D 아티스트들은 많은 시간을 들여 컴퓨터 그래픽 도구(CG)로 이미지를 수정하는 작업을 거친다. 이 때 합성 데이터로 생성된 대규모 이미지를 사용하면 완전히 사실적이지는 않을지라도 그래픽 아티스트의 작업을 일부분 자동화하기 때문에 이미지당 소요되는 비용이 적다.

AI기반 컴퓨터 비젼의 훈련에 사용된 합성 데이터 (출처: Datagen)
데이터 프라이버시의 보호
데이터가 증가하면 데이터 거버넌스에 대한 책임도 커지게 된다. 합성 데이터가 주목을 받게 된 것은 실제 데이터 확보 과정이 복잡하고 비용이 많이 들고 개인정보 유출에 대한 우려가 커졌기 때문이다. 실제 데이터를 그대로 사용하면 개인의 민감한 정보(인종, 성별, 정치 성향, 질병 기록)이 드러날 수 있지만 합성 데이터의 경우는 그 확률이 줄어든다.

합성 데이터가 잠재적으로 개인정보를 보호하는 이유는 원본 데이터 세트의 통계적 변수 분포와 상관관계 등을 모방하지만 정확한 데이터 포인트(data points)를 포함하지 않기 때문이다. 만일 어떤 데이터가 누구의 것인지를 추적할 수 없다면 그 데이터는 법적으로 개인 정보도 아니다.

MIT의 Data-To-AI 연구그룹이 개발한 '합성 데이터 보관소'(SDV: Synthetic Data Vault)는 개인 정보를 노출할 수 있다는 우려 때문에 데이터를 충분히 활용하지 못하는 문제에 대한 해결책을 제시했다. SDV는 민감 데이터가 포함된 원본 데이터 자체가 아니라 그것과 형식 및 구조가 유사한 합성 데이터를 생성하여 데이터의 활용성을 높여준다.

SDV파이썬 라이브러리가 합성 데이터 세트를 모델링하는 과정 (출처)
신경망 모델에 대한 개인정보 침해 공격이 늘어나자 합성 데이터의 유용성을 유지하면서도 개인 정보의 재식별을 막는 메커니즘을 찾으려는 연구가 계속되고 있다. 예컨대, 신시아 드워크(Cynthia Dwork)가 개발한 차분 프라이버시(differential privacy)란 무작위적으로 잡음을 추가하여 데이터를 변경하거나 수정하는 기법으로서 개인 데이터가 식별될 수 있는 위험을 낮추는 방법이다.

차분 프라이버시 보호를 만족하는 안전한 GAN 기반의 합성 데이터는 가장 이상적 데이터로 여겨진다. 미국 국립표준기술 연구소(NIST)는 합성 데이터 생성 알고리즘이 차등 정보보호 요건까지 충족하도록 하는 설계방법을 도전 과제로 내걸었다.

맹신은 위험하다
합성 데이터를 사용하면 민감한 개인 정보가 재식별(re-identify)될 가능성이 감소하지만 완벽하지는 않다. 따라서 합성 데이터 솔루션이 민감한 정보의 노출을 완전히 방지하거나 재식별을 위한 공격에 면역적이라는 선전은 과장된 것이다.

미국의 건강보험 양도와 책임성 법률(HIPAA)에 따른 개인정보 비식별화 가이드라인은 데이터 간의 구분가능성(distinguishability)이 0.04% 미만일 때 정보 주체가 재식별되는 리스크가 없다고 본다. 그렇지만 현재 합성 데이터 생성용 알고리즘에서 정보 재식별율은 10%를 상회한다. 예컨대 두 개 이상의 데이터 항목을 조합하거나 데이터 분포가 편중되어 있다면 그 민감한 정보가 누구의 것인지 식별이 가능할 수 있다.

한편, 합성 데이터를 이용하면 데이터에 내재된 편향의 문제가 완전히 해소될 수 있다는 오해도 종종 발견된다. 그러나 합성 데이터로 훈련한 AI모델이 편향성이 없거나 윤리적이라고 단정하기는 어렵다.

합성 데이터를 이용하면 균형이 맞지 않는 데이터 세트와 관련된 편향을 감소시키지만 합성 데이터 생성의 기초가 된 실제 데이터에 숨겨진 편향을 그대로 반영할 수도 있다. 다시 말해서 합성 데이터는 원본 데이터의 통계적 변수 분포와 상관관계 등을 모방하므로 합성 데이터 품질을 무조건적으로 믿는다면 위험할 수 있다. 예컨대 어떤 AI모델에는 흑인 소비자의 데이터는 없고 백인 위주의 데이터로만 학습했거나, 현실 세계의 암묵적 편향이 데이터 세트에 장기간 축적되었던 문제가 발견되기도 한다.

그럼에도 불구하고 합성 데이터는 현재의 AI 모델의 훈련 방식을 바꿔 놓을 것으로 보인다. 합성 데이터는 데이터 부족을 해결하고 민감한 데이터의 사용과 관련된 제약을 최소화하고 기

계학습 모델의 정확도를 높일 수 있다는 점에서 AI의 미래로 여겨지고 있다. 2024년에는 AI 연구개발에 필요한 데이터의 60% 가량을 합성 데이터가 담당하게 될 것이다.

* 최은창은 MIT 테크놀로지리뷰 한국판 편집장이며, 옥스퍼드대 법대 방문학자, 과학기술정책연구원(STEPI) 펠로우, 예일대 로스쿨의 정보사회프로젝트(Yale ISP) 펠로우로 연구했다. 저서로 《레이어 모델》,《가짜뉴스의 고고학》, 공저로 《인공지능 윤리와 거버넌스》,《인공지능 권력변환과 세계정치》, 《20개의 핵심 개념으로 읽는 디지털 기술사회》 등이 있다.

합성 데이터의 시대가 오고 있다

신뢰도 높은 AI 시스템을 구축하려면 양질의 데이터가 필요하지만 AI 모델의 훈련에 필요한 데이터를 구하기는 쉽지 않다. 원본 데이터의 통계적 변수 분포와 상관관계 등을 모방한 합성 데이터(재현 데이터)는 고질적인 데이터 병목현상을 해소할 수 있다.
2022년 10월 24일

데이터 부족의 문제

AI 개발자들은 몇 가지 골치 아픈 이슈들에 직면해 있다. 우선, AI 개발 과정에서 기업들은 데이터를 절실히 필요로 한다. 올바른 데이터 구하기는 강력한 AI를 구축하는 데 가장 중요하면서도 가장 어려운 부분이다. 예컨대, 의료용 AI 개발자가 고품질의 병변 데이터를 구하기는 어렵다. 이러한 데이터 공급의 한계는 AI의 발전 속도를 느려지게 만드는 요인이다.
둘째, 데이터 품질이 낮거나 데이터 세트에서 개인 정보가 노출되는 문제가 종종 발생한다. 데이터 품질의 문제는 AI 모델의 판단이 편향되거나 공정(fair)하지 않을 수 있다는 불신으로 번지게 된다. “쓰레기 데이터를 넣는다면 쓰레기가 나온다(garbage in, garbage out)”는 격언은 실무에서 여전히 유효하다. 질 낮은 데이터는 AI 모델의 연산을 거친 결과값을 신뢰할 수 없게 만든다.
셋째, AI 모델에 데이터를 공급할 때 데이터의 원본에서 개인 정보를 제거하고 사회적 불평등 논란을 미리 방지해야 한다. 정확한 예측 결과를 제공하려면 데이터 세트가 편향되지 않아야 하고 강화된 개인 정보보호 규정까지 준수해야 한다. 데이터를 구하기가 어려워지자 많은 기업들은 합성 데이터(synthetic data)에 주목하기 시작했다. 합성 데이터를 사용하면 훨씬 빠르고 적은 비용으로 AI 모델의 훈련 데이터를 확보할 수 있다. 오늘날 세계에서 가장 귀중한 자원은 데이터인데, 이 데이터를 무한한 양으로 저렴하고 빠르게 생산할 수 있는 방법이 있다면 기업의 입장에서는 관심을 가질 수밖에 없다.

합성 데이터는 어떻게 생성되나?

MIT 테크놀로지 리뷰는 2022년 중의 하나로 합성 데이터를 선정했다. 합성 데이터는 ‘재현 데이터’로도 불리는데 실제 데이터 세트에 존재하는 통계 패턴을 모방한 데이터()를 의미한다.

합성 데이터에 대하여 유럽 데이터 보호 감독기구(EDPS)는 “원래 데이터 소스를 가져와서 유사한 통계 속성을 가진 새로운 인공 데이터를 생성”하는 것이라고 정의한다. 간단히 말해, 합성 데이터는 데이터의 통계적 특성을 모방하여 만들어진 인공적으로 만들어진 가짜 데이터이다.

그렇다면 합성 데이터는 어떻게 만들어지는 것일까? 현실 세계의 사건들을 실제로 수집하지 않고 컴퓨터 시뮬레이션이나 알고리즘이 생성한다. 즉, 소량으로 수집된 원본 데이터 세트를 샘플로 삼아서 그 통계적 특성을 모방하여 인위적으로 만들어진다.

신경망을 이용한 합성 데이터의 생성은 생성적 적대 신경망(GAN:Generative Adversarial Networks)에서 힌트를 얻었다. 캐나다 몬트리올대에서 요슈아 벤지오의 지도 아래 박사과정을 마친 이언 굿펠로우(Ian Goodfellow)는 동료들과 이미지를 정밀하게 그리는 기계를 개발하려는 아이디어에 대해 이야기를 나누었다. 그는 적대적 양방향의 신경망들이 서로 경쟁하면서 가짜 이미지를 사실적으로 정확하게 생성하는 방법을 생각해 냈다. 합성 데이터를 인위적으로 생성하는 방법은 생성적 적대 신경망의 작동방식으로부터 자연스럽게 얻어진 것이다.

합성 데이터의 장점은 무엇일까?

AI 모델을 훈련시킬 때는 정확히 레이블(label)된 풍부한 데이터 세트가 필요하다. 더욱 다양한 데이터로 훈련한다면 더 높은 정확도를 달성할 수 있지만 수백만 개의 대규모 데이터를 수집하고 레이블을 지정하는 작업에는 막대한 시간과 비용이 필요하다. 게다가, 실제 데이터(real-world data)는 AI 모델의 훈련에 적합하지 않은 경우도 많다. 직접 관찰하여 얻은 데이터보다 합성 데이터가 가치 있다니 언뜻 보기에 모순적으로 들릴 것이다.

실제 데이터는 물론 좋은 통찰력을 제공하지만, 실제 데이터는 우연에 좌우되는 경우가 많고 현실 세계에서 가능한 모든 조건이나 사건의 순열을 포함하지 않는다. 게다가 실제 데이터는 개인정보 보호 규정으로 인해 데이터 전처리 과정(preprocessing)에 비용이 많이 들고 엉망인 상태(messy)이거나 오염된 경우가 많다.

실제 데이터에는 부정확한 요소들과 편향(bias)까지도 포함되어 있기 때문에 데이터 정제(data cleansing) 과정을 거치지 않으면 신경망에 오히려 악영향을 미칠 수가 있다. 즉, 실제 데이터를 수집한 이후에는 데이터 전처리 과정을 거쳐서 개인 정보를 제거하고, 오류를 걸러내고 서로 다른 데이터 형식들도 통일해야만 한다. 이 과정은 번거롭고 비용을 증가시킨다.

저명한 벤처 투자가 롭 토우()는 이 문제를 정확히 지적했다. AI 모델을 구축하는 많은 기업들이 이미지 레이블링 작업에 해마다 수천만 달러가 넘는 돈을 쏟아붓고 있지만 그것이 실제로100% 정확하리라는 보장이 없다고 말한다. 반면, 합성 데이터를 사용하여 AI모델을 훈련시킨다면 더욱 더 균일한 데이터 형식과 레이블을 유지할 수 있으므로 실제 데이터의

효과적인 보완책이나 대안으로 여겨지고 있다.

훈련용 데이터의 공급

합성 데이터는 레이블이 지정된 훈련 데이터를 거의 무제한으로 생성하여 심층 신경망(DNN, Deep Neural Network)에 공급할 수 있다. 또한 합성 데이터를 사용한 심층 신경망 훈련은 실제 데이터로 훈련시키는 경우에 비해 더 적은 비용이 들어간다. 합성 데이터의 확장성도 매력적인 요소이다. 머신러닝 모델을 훈련하고 테스트하는 데 필요한 데이터를 확보하는 일은 까다로운 과정이 요구되지만 합성 데이터는 생성과 사용이 더 간단하다.

예컨대, 간 병변 데이터 이미지만가 적은 분량만 있다면 질병 진단용 AI 모델을 훈련시켜도 예측의 정확도가 떨어지는 문제가 있었다. 그런데 의료용 합성 데이터를 AI 모델의 훈련에 추가적으로 사용하자 진단용 신경망의 분류 성능이 확연히 되었다. GAN을 사용하여 실제 데이터와 같은 가상의 의료용 영상 데이터를 생성하게 되자 환자들이 질병의 진단을 위해 방사선 검사를 받을 필요성도 크게 줄어들었다.

의료용 데이터는 민감하므로 신원이 노출되어서는 안되지만 합성 데이터는 개인정보 보호와 데이터 유용한 활용 사이에서 균형을 찾을 수 있는 방법이다. 이러한 맥락에서, 옥스포드대 바이오의료 공학과의 엘리슨 노블(Alison Nobel)은 합성 데이터가 민감한 의료 정보의 공정이용(fair use)을 확대하는 역할을 할 것이라고 한다.

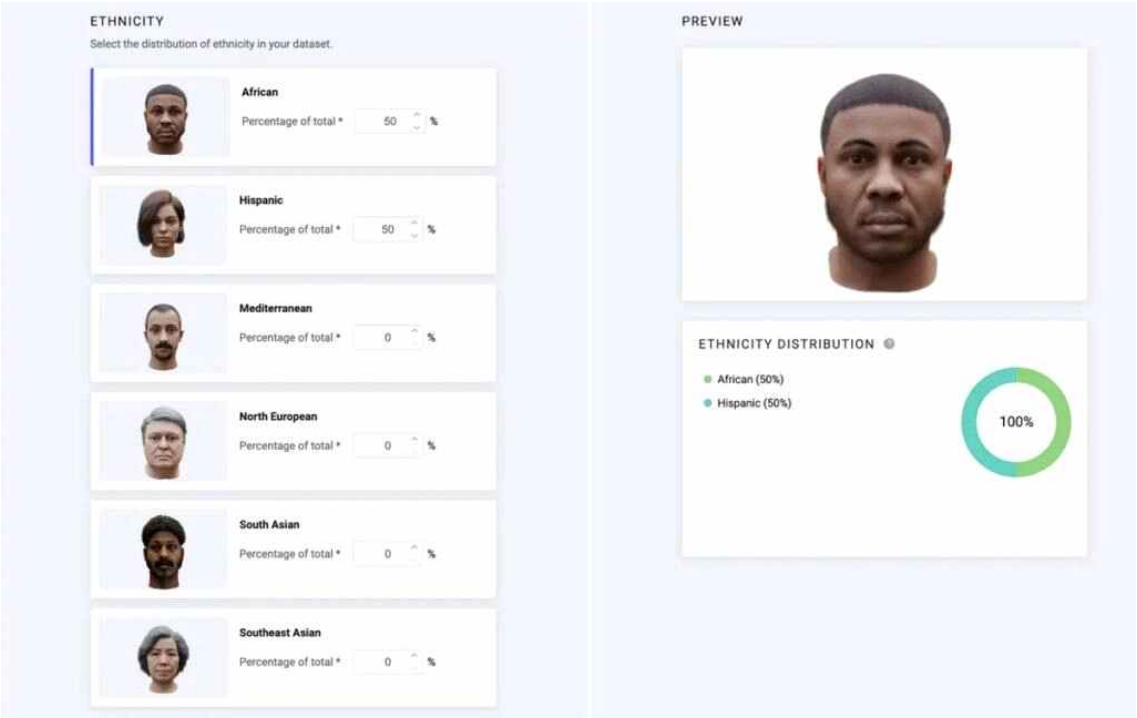
합성 데이터 스타트업들

최근에는 개인정보 유출에 대한 우려가 커졌고 데이터 기반 솔루션 수요도 늘어났다. 이런 흐름은 합성 데이터의 활용도를 높이는 배경이 된다. 합성 데이터는 원래 데이터 세트의 통계적 특성 또는 분포를 유지하면서도 익명성을 유지시켜 준다.

합성 데이터를 다양한 용도를 위해서 제공하는 스타트업들은 2021년부터 주목을 끌고 있다. 샌프란시스코에 위치한 '신세시스 AI'(는 데이터 다양성(data diversity)에 초점을 두고 주문형 합성 데이터를 기업들에게 제공한다. 수만 명이 넘는 특이한 사람들, 개체들, 환경을 결합하여 거의 무한에 가까운 데이터 가변성을 생성한다. 컴퓨터 비전용 합성 데이터는 증강 현실(AR)과 가상 현실(VR)에서 디지털 아바타의 자세와 동작 인식 등에도 사용될 전망이다.

이스라엘 기업 데이터젠()은 합성 데이터가 점차 컴퓨터 비전 시스템의 핵심을 담당할 것으로 내다보고 있다. 이 업체는 AI 개발자들에게 합성 데이터 생성 플랫폼을 통해 컴퓨터 비전 모델의 훈련을 돕는다. 동작 기반 시퀀스의 합성 데이터 세트는 객체 감지 및 식별에 사용된다. 딥러닝 기반의 컴퓨터 비 모델이 피부색에 따라서 안면 인식률에 큰 차이를 보인다면 논란이 불거지게 된다. 그런데 다양한 인종의 얼굴들을 포함한 합성 데이터로 훈련시킨다면 피부색에 따른 인식률 차이는 크게 줄어들게 된다

한편, 엔터테인먼트 콘텐츠의 제작에도 합성 데이터가 사용된다. 3D 아티스트들은 많은 시간을 들여 컴퓨터 그래픽 도구(CG)로 이미지를 수정하는 작업을 거친다. 이 때 합성 데이터로 생성된 대규모 이미지를 사용하면 완전히 사실적이지는 않을지라도 그래픽 아티스트의 작업을 일부분 자동화하기 때문에 이미지당 소요되는 비용이 적다.



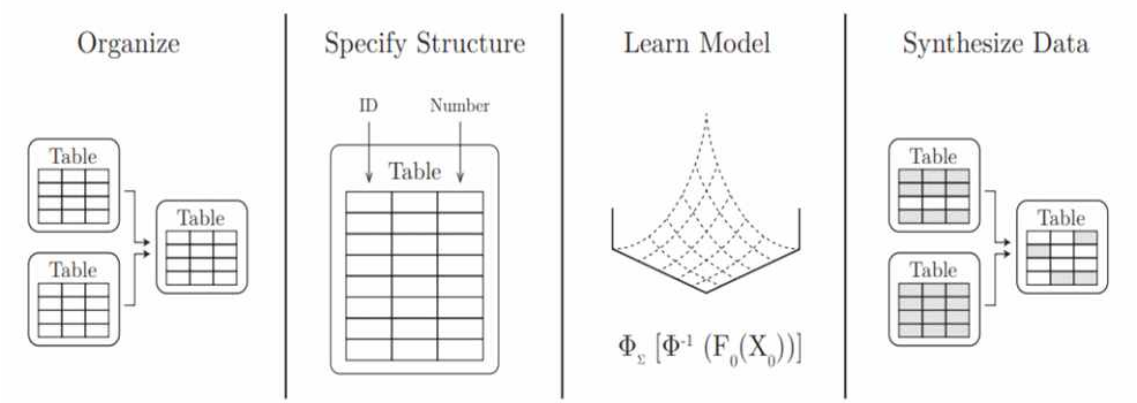
AI기반 컴퓨터 비젼의 훈련에 사용된 합성 데이터 (출처: Datagen)

데이터 프라이버시의 보호

데이터가 증가하면 데이터 거버넌스에 대한 책임도 커지게 된다. 합성 데이터가 주목을 받게 된 것은 실제 데이터 확보 과정이 복잡하고 비용이 많이 들고 개인정보 유출에 대한 우려가 커졌기 때문이다. 실제 데이터를 그대로 사용하면 개인의 민감한 정보(인종, 성별, 정치 성향, 질병 기록)이 드러날 수 있지만 합성 데이터의 경우는 그 확률이 줄어든다.

합성 데이터가 잠재적으로 개인정보를 보호하는 이유는 원본 데이터 세트의 통계적 변수 분포와 상관관계 등을 모방하지만 정확한 데이터 포인트(data points)를 포함하지 않기 때문이다. 만일 어떤 데이터가 누구의 것인지를 추적할 수 없다면 그 데이터는 법적으로 개인정보도 아니다.

MIT의 Data-To-AI 연구그룹이 개발한 '합성 데이터 보관소'(.)는 개인 정보를 노출할 수 있다는 우려 때문에 데이터를 충분히 활용하지 못하는 문제에 대한 해결책을 제시했다. SDV는 민감 데이터가 포함된 원본 데이터 자체가 아니라 그것과 형식 및 구조가 유사한 합성 데이터를 생성하여 데이터의 활용성을 높여준다.



SDV파이썬 라이브러리가 합성 데이터 세트를 모델링하는 과정 ()

신경망 모델에 대한 개인정보 침해 공격이 늘어나자 합성 데이터의 유용성을 유지하면서도 개인 정보의 재식별을 막는 메커니즘을 찾으려는 연구가 계속되고 있다. 예컨대, 신시아 드워크(Cynthia Dwork)가 개발한 차분 프라이버시(란 무작위적으로 잡음을 추가하여 데이터를 변경하거나 수정하는 기법으로서 개인 데이터가 식별될 수 있는 위험을 낮추는 방법이다.

차분 프라이버시 보호를 만족하는 안전한 GAN 기반의 합성 데이터는 가장 이상적 데이터로 여겨진다. 미국 국립표준기술 연구소(NIST)는 합성 데이터 생성 알고리즘이 차등 정보 보호 요건까지 충족하도록 하는 설계방법을 도전 과제로 내걸었다.

맹신은 위험하다

합성 데이터를 사용하면 민감한 개인 정보가 재식별(re-identify)될 가능성이 감소하지만 완벽하지는 않다. 따라서 합성 데이터 솔루션이 민감한 정보의 노출을 완전히 방지하거나 재식별을 위한 공격에 면역적이라는 선전은 과장된 것이다.

미국의 건강보험 양도와 책임성 법률(HIPAA)에 따른 개인정보 비식별화 가이드라인은 데이터 간의 구분가능성(distinguishability)이 0.04% 미만일 때 정보 주체가 재식별되는 리스크가 없다고 본다. 그렇지만 현재 합성 데이터 생성용 알고리즘에서 정보 재식별율은 10%를 상회한다. 예컨대 두 개 이상의 데이터 항목을 조합하거나 데이터 분포가 편중되어 있다면 그 민감한 정보가 누구의 것인지 식별이 가능할 수 있다.

한편, 합성 데이터를 이용하면 데이터에 내재된 편향의 문제가 완전히 해소될 수 있다는 오해도 종종 발견된다. 그러나 합성 데이터로 훈련한 AI모델이 편향성이 없거나 윤리적이라고 단정하기는 어렵다.

합성 데이터를 이용하면 균형이 맞지 않는 데이터 세트와 관련된 편향을 감소시키지만 합성 데이터 생성의 기초가 된 실제 데이터에 숨겨진 편향을 그대로 반영할 수도 있다. 다시 말해서 합성 데이터는 원본 데이터의 통계적 변수 분포와 상관관계 등을 모방하므로 합성 데

이터 품질을 무조건적으로 믿는다면 위험할 수 있다. 예컨대 어떤 AI모델에는 흑인 소비자의 데이터는 없고 백인 위주의 데이터로만 학습했거나, 현실 세계의 암묵적 편향이 데이터 세트에 장기간 축적되었던 문제가 발견되기도 한다.

그럼에도 불구하고 합성 데이터는 현재의 AI 모델의 훈련 방식을 바꿔 놓을 것으로 보인다. 합성 데이터는 데이터 부족을 해결하고 민감한 데이터의 사용과 관련된 제약을 최소화하고 기계학습 모델의 정확도를 높일 수 있다는 점에서 AI의 미래로 여겨지고 있다. 2024년에는 AI 연구개발에 필요한 데이터의 가량을 합성 데이터가 담당하게 될 것이다.

** 최은창은 MIT 테크놀로지리뷰 한국판 편집장이며, 옥스퍼드대 법대 방문학자, 과학기술정책연구원(STEPI) 펠로우, 예일대 로스쿨의 정보사회프로젝트(Yale ISP) 펠로우로 연구했다. 저서로 《레이어 모델》, 《가짜뉴스의 고고학》, 공저로 《인공지능 윤리와 거버넌스》, 《인공지능 권력변환과 세계정치》, 《20개의 핵심 개념으로 읽는 디지털 기술사회》 등이 있다.*