2022-10호

AI TREND WATCH

2022. 12. 30.

합성 데이터(Synthetic data)의 부상

정보통신정책연구원 김민진 전문연구원

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

합성 데이터(Synthetic data)의 부상

정보통신정책연구원 김민진 전문연구원

개 요

◈ **머신러닝에서 불충분한 데이터 이슈를 해소하는 대안으로 합성 데이터 (Synthetic data)가 부상하고 있음**

▶ 인공지능 기술 발전과 함께 데이터의 중요성이 증가하고 있으나, 개인정보 위험이 낮으면서 충분한 양의 고품질의 데이터를 확보하는 데에 어려움이 존재함

▶ 데이터 이슈 해소를 위해 합성 데이터가 대안의 일환으로 주목받고 있음

※ 합성 데이터: 실제 데이터세트의 통계 패턴을 모방하여 인공적으로 만들어진 데이터

◈ **본 고에서는 합성 데이터의 특징 및 산업 전망, 활용사례 및 장단점을 소개한 후 산업 및 정책수립의 시사점을 제시하고자 함**

주요내용

◈ **인공지능 기술에 있어 데이터가 중요하나, 여전히 데이터 확보에 어려움이 존재함**

▶ 데이터는 원유에 비유(Data is a new oil)될 정도로 인공지능 기술을 통한 혁신을 달성하기 위해서는 데이터가 중요

▶ 그러나 실제 활용 가능한 빅데이터를 확보하는 데에 제약요인이 존재함

– (데이터 접근성) 개인정보 보호 등의 이유로 데이터 접근 제한

– (데이터 품질) 데이터 정제(노이즈 제거)에 시간과 노력이 소요되며, 정확성, 완전성, 일관성, 적시성, 유효성 측면에서 고품질의 데이터 수집에 어려움

▶ 불충분한 데이터는 인공지능 예측 모델의 부정확성*을 초래할 수 있음

* (과적합, overfitting) 학습 오류가 적고 테스트 오류가 높아 현실 세계에서는 잘 작동하지 않는 경우

* (과소적합, underfitting) 지나치게 단순한 모델을 구축함에 따라 모델이 데이터의 관계를 캡처할 수 없어 예측력이 없는 경우

◈ **합성 데이터(Synthetic Data)는 개인정보를 보호하면서 불충분한 데이터로 인한 인공지능 모델 성능 문제를 극복하는 대안으로 부상**

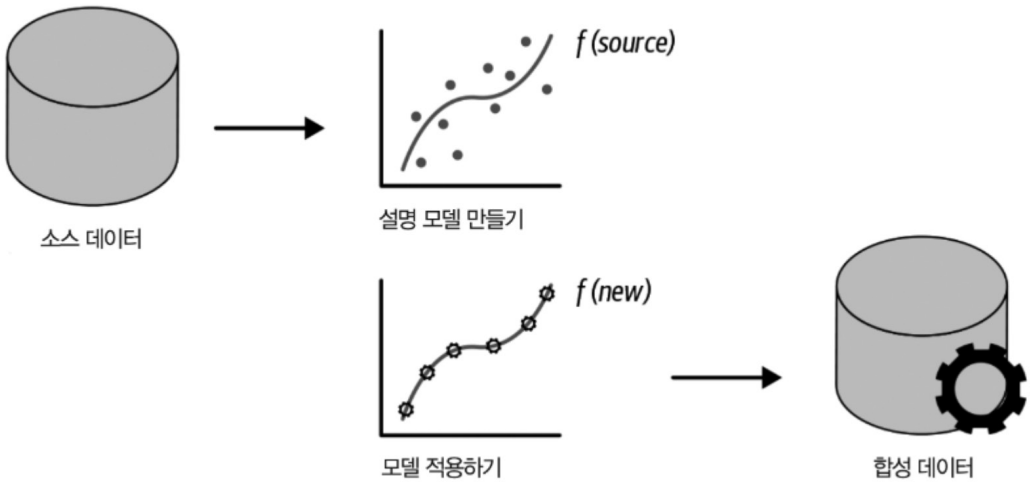
▶ (정의) 실제 데이터 세트에 존재하는 통계패턴을 모방하여 인공적으로 만들어진 가짜 데이터

- 실제 세계에서 수집되거나 측정되는 것이 아니라 디지털 세계에서 생성하는 것으로, 수학적으로 또는 통계적으로 실제 데이터를 반영
- 유럽데이터 보호 감독기구(EDPS)는 “원래 데이터 소스를 가져와 유사한 통계 속성을 가진 새로운 인공데이터를 생성하는 것”으로 정의(최은창, 2022)

▶ (생성 방법) 실제 데이터 없이 합성하는 방법과 실제 데이터로 합성하는 방법으로 구분되며 (한빛미디어, 2021) 실제 데이터 기반 합성 데이터 생성에도 다양한 세부 방법이 존재(James et al., 2021)

- (실제 데이터 없이 합성) 기존 모델(통계적 모델, 설문조사나 그 외 데이터 수집 메커니즘을 통해 개발된 것 등)이나 분석가의 지식을 이용해 생성
- (실제 데이터 기반 합성) 데이터를 설명하는 생성 모델을 사용해 합성 데이터를 생성

[그림 1] 실제 데이터를 기반한 데이터합성 프로세스



자료: 한빛미디어(2021)

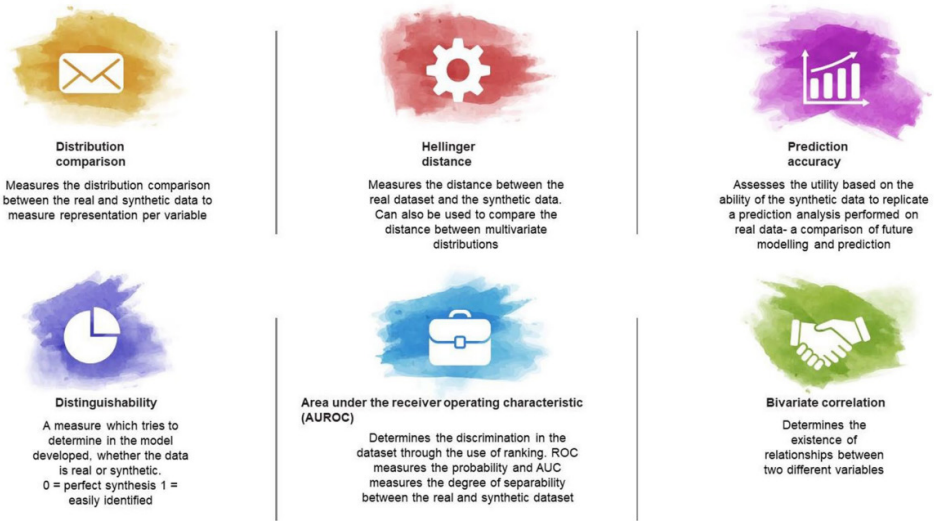
- (임베딩, Embedding) 두 개의 신경망(인코더와 디코더)을 사용하여 데이터를 생성하는 방법. 가변 자동 인코더가 원본 데이터를 압축하고 디코더로 보내며, 디코더는 원본데이터 세트를 나타내는 출력을 생성함. 시스템 학습은 입력 데이터와 출력 데이터 사이의 상관관계를 최대화 하는 것을 포함함
- (생성적 적대 신경망, GAN) 판별자는 생성자가 생성한 레코드를 실제 데이터와 구별하려는 시도를 하면서 학습하고 성향 점수를 제공

- (순차적 합성, Sequential synthesis) 변수별로 데이터 세트 변수를 합성하는 방법

▶ (합성 데이터의 품질 평가) 데이터 효용을 측정하고 원 데이터 세트와의 유사성을 확인하는 대표적인 합성 데이터 품질 지표는 다음과 같음(James et al., 2021)

- (분산 비교, Distribution comparison) 실제 데이터와 합성 데이터 간의 분산을 비교/측정하여 변수별 대표성 비교
- (헬링거 거리,Hellinger distance) 실제 데이터 세트와 합성 데이터 사이의 거리 측정
- (예측 정확도, Prediction accuracy) 실제 데이터에서 수행된 예측 분석과 합성 데이터에 기반한 예측 분석 성능을 비교하여 합성 데이터의 모방 가능성 확인
- (식별력, Distinguishability) 개발된 모델에서 데이터가 실제인지 합성인지를 결정하는 척도 활용(0 =완벽한 합성, 1= 쉽게 식별 가능)
- (AUROC) 다양한 임계값에서 합성 데이터 세트와 실제 데이터 세트를 분류하는 성능을 측정
- (이변량 상관관계, Bivariate correlation) 서로 다른 두 변수 간 관계 확인

[그림 2] 합성 데이터 품질 지표



자료: James et al. (2021), p. 9

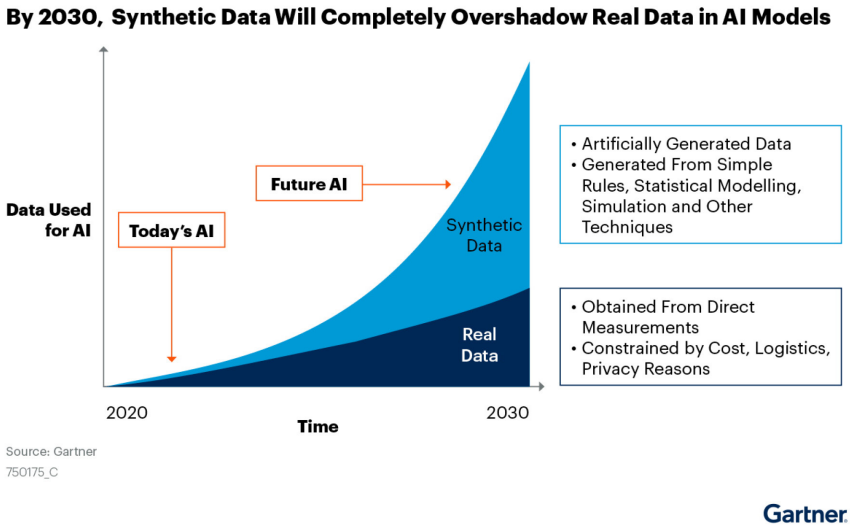
◈ **합성 데이터는 주요한 학습용 데이터로서 다양한 분야에서 활용이 기대되며, 관련 시장도 크게 성장할 것으로 예상됨**

▶ (전망) 합성 데이터는 인공지능에서 사용되는 주요한 데이터 형태가 될 수 있음

- MIT 테크놀로지 리뷰는 합성 데이터를 2022년 10대 미래 기술 중의 하나로 선정(최은창, 2022)

- 2024년까지 합성 데이터가 인공지능 학습용 데이터의 60%를 차지, 2030년까지 실제 데이터의 대부분을 대체할 것으로 예상(Gartner, 2021)

[그림 3] 인공지능 모델에서 합성 데이터와 실제 데이터의 활용 전망



자료: Gartner(2021)

▶ (합성 데이터 산업 현황) 합성 데이터를 제공하는 스타트업이 주목을 끌고 있으며, 글로벌 플랫폼 기업은 합성데이터 기업과 M&A 시도

- (글로벌 합성 데이터 기업) 신세시스 AI(Synthesis AI), 데이터젠(Datagen) 등의 스타트업이 주목을 받고 있으며, 글로벌 플랫폼 기업의 관련 기업 M&A가 이루어짐

〈표 1〉 글로벌 합성 데이터 스타트업 사례

기업명	특징(주요 서비스 모델)	비고
Synthesis AI	- 데이터 다양성에 초점을 두고 주문형 합성데이터 제공	
Datagen	- 합성 데이터 생성 플랫폼을 제공하여 컴퓨터 비전 모델의 훈련 지원	
AI, Reverie	- 국방, 소매, 산업 등 합성데이터 생성 플랫폼 - 합성 데이터 기반 비디오//이미지 생성 플랫폼	메타(전 페이스북)에 2021년 인수
Caper	- 합성데이터로 상품 인식 기능을 높여 스마트카트 셀프 결제 서비스 개선	인스타카트에 2021년 인수(USD 350M)
Tonic.ai	- 합성 데이터 생성 도구 제공	시리즈 B, 2021년 (USD 45M)
Gretel Labs	- 합성 데이터로 개인정보 침해 없이 통계활동 지원	시리즈 B, 2021년 (USD 68M)

자료: 최은창(2022) ; 김윤진(2022.5)를 정리함

– (국내 합성 데이터 현황) 국내 AI 합성 데이터 생성 시장 규모는 2024년까지 약 5,752억 원 규모로 성장할 전망

〈표 2〉 국내 AI 학습 데이터 생성 시장규모

(단위: 억 원, %)

2018	2019	2020	2022	2024	CAGR
1629	2010	2481	3061	5752	23.4

자료: 중소벤처기업부 외(2021)

◈ 산업별, 업무별 합성데이터 활용사례가 존재

▶ (산업별 합성 데이터 활용 사례) 합성 데이터는 의료, 보험, 금융 분야 등에 활용되어 연구 향상, 서비스 고도화, 고객 정보 보호 효과를 기대

– (의료 및 제약 분야) 실제 데이터를 사용할 수 없거나 데이터가 부족한 경우 합성 데이터를 사용 가능

- 합성 데이터를 통해 의료 데이터 전문가는 환자 기밀을 유지하면서 기록 데이터의 내·외부에서 사용 가능
- 실제 데이터가 아직 존재하지 않을 때 향후 연구 및 테스트에도 사용 가능

– (보험 분야) 서비스 개선에 있어 청구 데이터, 판매 및 이탈 데이터, 시장 및 설문 조사 데이터 등의 합성 데이터 사용 가능

- 합성 데이터는 부서간 실시간 정보 교환에 용이하여 고객 여정 개선, 리스크 관리, 언더라이팅 정확도 향상 등에 활용

– (금융 분야) 데이터 프라이버시를 보호하면서 사기탐지 고도화 등에 합성 데이터 사용 가능

- 합성 사기 데이터를 사용하여 새로운 사기 탐지 방법 테스트 및 효과 평가
- 대부분의 고객 데이터가 비공개인 금융 분야에서 고객 행동을 이해하기 위해 합성 고객 거래 데이터를 사용 가능

〈표 3〉 산업별 합성데이터 활용 사례

분야	데이터 타입	활용 사례
금융 서비스	금융 데이터, 트레이딩 데이터, 고객 데이터 등	· 사기 식별, 고객 분석
보험	고객 데이터, 보험 기록	· 신제품 개발 및 테스트, 고객 경험 개선, 언더라이팅 정확도 향상 등 (스위스 Die Mobiliar)
헬스케어	환자 데이터, 의료 기록	· 의료 분석, 임상 시험
자동차 및 로봇 공학	위치 정보(도로, 차, 신호, 건물 등), 이미지 정보(사고, 위험 상황 등)	· 자율주행차, 자율 로봇 (테슬라 AI 데이에 합성 데이터 사용 공개)

자료: 최은창(2022); Cem(2021); Fdal(2022)을 정리함

▶(업무별 합성 데이터 활용 사례) 기계 학습, 내부 소프트웨어 테스트, 교육, 훈련 및 해커톤, 데이터 보존, 공급업체 평가 및 타사 서비스와 데이터 공유, 내부 2차 사용, 외부 공유 등에 사용 가능

- (머신러닝) 기계학습 기술 평가 및 비교, 데이터 증대, 사이버 공격으로 인한 학습 데이터의 복구(재식별) 위험 감소
  - 작은 테스트 데이터에서 모델을 평가할 때 발생할 수 있는 샘플링 변동성을 제거
  - 데이터 세트 내에서 과소 표현된 하위 모집단을 강화할 때 합성 데이터를 추가하면 포괄적이고 공정한 모델 생성 가능
  - 사이버 공격에 의한 학습 데이터의 개인 정보 재식별 위험 감소 가능
- (내부 소프트웨어 테스트) 개인 데이터 사용 없이 현실적인 개별 고객/환자 수준의 데이터 테스트 가능
- (교육, 훈련, 해커톤) 개인 데이터 처리 방법을 이해해야 하는 직원에게 개인 정보 접속 없이도 효과적인 교육 도구로 활용 가능
- (데이터 보존) 데이터 유용성이 높고 재사용이 증가하는 경우 개인정보 보호 문제 없이 데이터 유용성을 유지하기 위해 합성 기술 적용 가능
- (공급업체 평가) 공급업체 및 타사 서비스와 데이터 공유 시 데이터 제공 프로세스를 가속화하고 개인 데이터 사용에 따른 불필요한 작업을 피할 수 있음
- (내부 2차 이용) 기본 목적 이외 추가 연구를 수행하기 용이함
- (외부 공유) 외부 데이터 접근을 가속화할 수 있음

〈표 4〉 업무별 합성데이터 활용 사례별 재무, 기술 및 조직, 평판 효과

활용 사례		재무 효과	기술적/조직적 효과	평판 효과
머신러닝	머신러닝 기술 평가 및 비교	비용 수반 없이 빅데이터 생성	학습, 검증, 테스트에 유용	프라이버시와 AI 염려 제거
	데이터 증강		모델 성능 향상	
	프라이버시 공격 방어	탐색 강화	학습, 검증, 테스트에 유용	프라이버시와 AI 염려 제거
내부 소프트웨어 테스트		sub-standard 데이터를 사용하여 불량률 감소	소프트웨어 엔지니어링, 신뢰성, 테스트에 유용	프라이버시 염려 제거
교육, 훈련, 해커톤		인재 유지	교육, 개발, 문제 해결에 유용	개인 데이터 처리 미흡 회피
데이터 보존		고비용의 개인 데이터 스토리지와 공급 필요성 감소	합성 데이터 생성을 위한 프로세스와 인재 필요	
공급업체 평가			합성 데이터 공유 유용	
내부 2차 이용		긴 협상을 줄이고 탐색 강화	데이터를 활용한 2차 분석, 탐색, 탐구 가능	개인 데이터 처리에 있어서 좋은 사례로 기록
외부 공유	규제 문제 완화		관할권 전반에 걸쳐 잠재적으로 쉽게 공유 가능	개인 데이터 처리에 있어서 좋은 사례로 기록
	데이터 접근성 강화		빠른 데이터 공유 가능	프라이버시 염려 제거

자료: James et al. (2021), p. 4

◈ **데이터 합성은 프라이버시 이슈에서 비교적 자유로운 많은 양의 데이터를 효율적으로 생성 가능하게 하며, 합성 데이터는 인공지능 모델 성능을 향상하는 데에 기여**

▶ 효율적인 데이터 생성(유소영 외, 2022)

- 사람이 구축하는 실제 데이터는 수집과 라벨링에 장시간과 고비용이 소요되나, 합성 데이터는 컴퓨터 자동화 기술을 통해 대량의 데이터를 빠르고 저렴하게 생성할 수 있음

  * AI Reverie의 공동 설립자(Paul Walborsky)는 인공 데이터 생성이 수작업 라벨링 서비스의 약 100분의 1 정도의 비용이 발생한다고 주장

▶ 정보 보호와 데이터 활용 향상(최은창, 2022)

- 데이터는 일반적으로 개인의 동의하에 특정 용도로 수집되는데, 이를 다른 용도로 사용한다면 이는 이차 목적으로 간주하며, 개인 데이터를 이차 목적으로 사용할 때 법적 근거를 요구*(한빛미디어, 2021)

  * 미국의 의료보험 이전과 책임에 관한 법률(Health Insurance Portability and Accountability Act, HIPAA), 유럽의 일반 데이터 보호 규정(General Data Protection Regulation, GDPR) 등의 개인 정보 보호 규정은 개인 데이터를 이차 목적으로 사용할 때 법적 근거를 요구

- 합성 데이터는 식별 가능한 개인 데이터로 간주되지 않아* 개인 정보 보호 규정이 적용되지 않으며, 데이터를 이차 목적으로 사용하기 위한 추가 동의도 필요 없게 됨(한빛미디어, 2021)

  * 합성 데이터는 원본 데이터 세트의 통계적 변수 분포와 상관관계 등을 모방하지만, 정확한 데이터 포인트(Data point)를 포함하지 않아 데이터가 누구의 것인지 추적 불가

- 또한, 합성 데이터가 프라이버시 문제에서 벗어나게 되면 데이터 저장 제한 원칙* 제약에서도 자유로울 수 있음

  * GDPR에는 개인 데이터를 필요 이상으로 오래 보관해서는 안된다는 저장 제한 원칙이 존재

▶ 인공지능 모델 성능 향상(유소영 외, 2022)

- 충분한 양의 데이터와 다양성을 확보하며 Bais를 줄여 인공지능 모델의 강건성 향상 가능

  * 합성 데이터가 실제 개체, 이벤트 또는 사람을 기반으로 하는 데이터보다 AI 모델 교육에 더 좋을 수 있음(Tremblay et al., 2018)

◈ **그러나 합성 데이터에도 데이터 생성 방법 결정에 대한 전문성이 요구되며, 프라이버시 이슈, 데이터 편향에서 완전히 자유롭지 못하다는 점에서 신중한 접근이 필요**

▶ (시간 및 숙련성 필요) 합성 데이터를 생성하는 방법을 적용하는 데에도 많은 시간이 소요됨

- 데이터 생성 방법이 다양하며, 생성 방법에 따른 장단점이 존재하므로 적용 방법을 선택하는 데에 시간이 필요

▶ (프라이버시 이슈 해소 노력 필요) 민감한 개인 정보가 재식별될 가능성이 감소하지만 여전히 관련 이슈가 존재

- 미국의 의료보험 이전과 책임에 관한 법률(HIPAA) 데이터 간의 구분가능성이 0.04% 미만일 때 정보 주체가 재식별되는 리스크가 없다고 보지만, 현재 합성 데이터 생성 알고리듬에서 정보 재식별률은 10%를 상회(최은창, 2022)

- 따라서 개인을 재식별하지 못하도록 하기 위해서 비즈니스 프로세스, 개인정보 보호 규정에 대한 숙련된 전문가가 필요

▶ (데이터 편향 해소 노력 필요) 합성 데이터는 편향의 문제를 완전히 해소할 수 없음

- 합성 데이터 생성의 기초가 된 실제 데이터에 숨겨진 편향을 그대로 반영할 위험 존재

시사점

◈ **인공지능 기술 활용에 있어 개인 정보 보호 문제와 스몰 데이터 한계를 넘어서기 위한 다양한 방안이 고려되었음**

▶ 개인 정보 보호 문제를 해소 방안으로 동형암호(윤성욱, 2020), 차분 프라이버시(Jung & Park, 2018) 등이, 스몰 데이터 제약 극복 방안으로 전이 학습(Transfer learning), 앙상블 학습 등이 논의되어 왔음

◈ **합성 데이터는 개인 정보 보호 문제를 줄이면서 적은 노력과 비용으로 무제한 데이터를 생성할 수 있다는 장점을 가지고 중요한 인공지능 학습 데이터로 부상**

▶ 합성 데이터는 원래 데이터 소스와 유사한 통계 속성을 가진 새로운 인공데이터를 생성하는 것으로, 개인 정보 보호와 데이터 유용성을 모두 확보하는 데에 용이

▶ 인공지능 학습에의 주요 데이터로 논의되고 있으며, 관련 시장도 꾸준히 성장할 것으로 전망

- Gartnet는 인공지능 학습에 있어 합성 데이터가 실제 데이터 활용을 능가할 것으로 예상하였으며, 다양한 산업과 업무에서 합성 데이터 활용 가능성이 논의되고 있음
- 글로벌 플랫폼 기업과 합성데이터 기업간 M&A가 이루어지고 있으며, 국내에서도 합성 데이터 생성 시장이 연평균 23.4% 성장할 것으로 전망

◈ **그러나 현업에서 합성 데이터를 활용하기 위해서는 여전히 고려할 이슈가 존재**

- 데이터 생성시 전문가의 개입이 요구되며 프라이버시 이슈와 데이터 편향의 문제에서 자유로울 수 없음
- 따라서 산업별, 업무별로 합성 데이터의 필요성과 적합성에 대한 지속적 논의가 필요하며, 한계를 극복하기 위한 다각적 노력(데이터 생성 방법의 복합적 접근 등)이 필요
- 또한, 데이터의 효용을 적용하고 합성 데이터 세트와 원 데이터 세트 간의 유사성을 지속적으로 모니터링하여 데이터 유용성을 확보해야할 것

참고문헌

김윤진(2022.5), 더 싸게, 더 많이, 더 빨리 With AI 시대 앞당길 가짜의 힘, 동아비즈니스리뷰.

유소영, 이성희, 김은지, & 강남우. (2022). 가상 제품 개발과 메타버스를 위한 3D 합성 데이터 생성, 기계저널. 62(11), 32-37.

윤성욱(2020.10.31.), 인공지능의 발전에서 동형암호가 갖는 의미. AI Trend Watch, KISDI.

중소벤처기업부 외(2021), 중소기업 전략기술로드맵 2021-2023 인공지능.

최은창(2022.10.24.), 합성 데이터의 시대가 오고 있다, MIT Technology Review.

한빛미디어(2021.3.9.), 합성데이터란 무엇인가(www.hanbit.co.kr/media/)

Dilmegani, Cem(2021.4.30.), What is Data Augmentation? Techniques, Examples & Benefits. AIMultiple.

Dilmegani, Cem (2021.7.10), "Top 20 Synthetic Data Use Cases & Applications in 2022". AIMultiple.

Fdal, A. Omar (2022.5.5.), Synthetic Data: 4 Use Cases in Modern Enterprises, Dataversity.

Gartner(20211.6.24), Maverick Research: Forget About Your Real Data – Synthetic Data Is the Future of AI.

James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: exploring use cases to optimise data utility. Discover Artificial Intelligence, 1(1), 1-13.

Jung, K., &Park, S. (2018). 차분 프라이버시 기반 비식별화 기술에 대한 연구. Review of KIISC, 28(2), 61-77.

NVIDIA(2021. 6. 8.), What Is Synthetic Data?.

Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., & Birchfield, S. (2018). Deep object pose estimation for semantic robotic grasping of household objects. arXiv preprint arXiv:1809.10790.