차등 프라이버시 기반 이미지 생성 기술 동향의 차원적 관점을 통한 분석

이성연 [1], 조윤기 [1], 백윤흥 [1]
[1] 서울대학교 전기 · 정보공학부, 서울대학교 반도체 공동연구소

sylee@sor.snu.ac.kr, ygcho@sor.snu.ac.kr, ypaek@snu.ac.kr

Trends in Differential Privacy-based Synthetic Image Data Generation: A Dimensional Analysis

Sungyeon Lee[1], Yungi Cho[1], Yunheung Paek[1]
[1]Dept. of Electrical and Computer Engineering (ECE), Seoul National University
[1]Inter-University Semiconductor Research Center (ISRC), Seoul National University

요 약

차등 프라이버시를 적용한 이미지 데이터 생성 기술은 기계 학습 모델의 민감 데이터 유출 위험으로부터 효과적으로 보호해준다. 최근 차등 프라이버시 기반의 이미지 생성 기술은 단순히 프라이버시를 보장하는 수준을 넘어, 데이터 유용성과 성능 사이의 균형을 고려하는 방향으로 발전하고 있다. 본 연구는 차원적 관점을 적용하여 이러한 발전 과정을 분석하고, 기술 패러다임의 동향을 재정의함으로써 향후 연구 방향을 제시하고자 한다.

1. 서론

데이터 기반의 현대 사회에서 개인정보 보호는 중요한 과제 중 하나로 부각되고 있다. 대규모 데이터셋을 활용한 기계 학습 모델은 다양한 분야에서 혁신을 주도하고 있지만, 동시에 개인의 민감한 정보 유출 위험을 내포하고 있기도 하다. 기존의 익명화 기법은 재식별 공격(Re-Identification Attack)에 취약하며, 회원 추론(Membership Inference)이나 데이터 재구성(Data Reconstruction)과 같은 정교한 공격은 학습 데이터 속 개인정보를 직접적으로 드러낼 수 있다.[1] 이러한 배경 속에서, 차등 프라이버시(Differential Privacy, DP)는 데이터 분석 결과가 특정 개인의 데이터 포함 여부에 거의 영향을 받지 않도록 보장하는 엄격한 수학적 프레임워크를 제공하고, 프라이버시 보호의 황금 표준(gold standard)으로서 자리 잡고 있다.[2]

차등 프라이버시는 민감도(sensitivity)에 비례하는 랜덤 노이즈를 주입함으로써 데이터의 프라이버시를 보장하는 원리에 기반한다. 수식적으로는 아래와 같이 정의된다.

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S]$$

이 식에서 D와 D'은 단일 개인의 데이터만 다른 두 데이터셋이고, $\mathcal{M}$은 분석 메커니즘, S는 출력 집합, 그리고 ϵ은 프라이버시 예산(budget)을 각각 의미한다. 이 정의는 데이터셋에 특정 개인이 포함되었는지 여부가 기계 학습 모델 출력에 미치는 영향을 최소화하도록 보장한다.

그러나 이미지와 같은 고차원 데이터에 이 차등 프라이버시를 직접 적용할 경우, 이른바 차원의 저주(curse of dimensionality)라는 근본적 한계에 직면하게 된다.[3] 차원이 증가함에 따라 의미 있는 신호를 유지하면서 동시에 프라이버시 제약을 만족시키기 위해 필요한 노이즈의 양이 기하급수적으로 증가하며, 이는 결과적으로 생성된 이미지의 품질을 심각하게 저하시킨다.

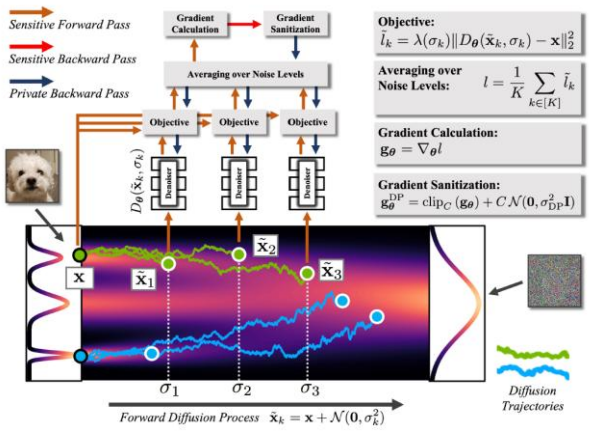
본 논문에서는 차등 프라이버시 기반 이미지 생성 기법의 발전사를 차원의 저주를 극복하기 위한 여정이라는 시각에서 재조명하고자 한다. 이러한 관점은 차등 프라이버시 메커니즘이 적용되는 위치에 따라 연구 패러다임의 전환 단계를 구분할 수 있게 해주며, 보다 낮은 차원의 표현 공간에서 효과적으로 프라이버시를 보장하는 문제를 탐구하는 과정으로 이해할 수 있다. 이를 통해 기술의 발전 경로를 체계적으로 정리하고, 향후 연구 방향을 제시하고자 한다.

2. 차원적 관점에서의 기술 패러다임 동향

2.1. 모델 단계 프라이버시 – 초고차원 공간

차등 프라이버시 이미지 생성의 초기 접근법은 생성 모델을 학습시키는 과정, 특히 모델의 가중치를 업데이트하는 그래디언트(gradient)에 직접 노이즈를 주입하는 방식이다. 이 패러다임의 핵심 알고리즘은 DP-SGD(Stochastic Gradient Descent)이다.[4] DP-SGD 는 각 학습 데이터가 모델 업데이트에 미치는 영향을 제한하기 위해 개별 샘플의 그래디언트 크기를 잘라내고(clip), 가우시안 노이즈를 추가하여 모델 파라미터를 업데이트한다. 따라서 노이즈가 추가되는 공간의 차원은 모델의 학습 가능한 파라미터 수와 동일하다. 예를 들어, 차등 프라이버시를 이용하여 수백만 개의 파라미터를 가진 확산 모델(diffusion model)을 학습시킨다면, 매 학습 단계마다 그에 상응하는 차원의 벡터에 노이즈를 추가해야 한다. 고차원에 걸친 노이즈를 분산시켜야 하므로, 각 파라미터 업데이트에 가해지는 노이즈의 영향이 매우 커진다. 이로 인해 모델 학습이 불안정해지고 수렴이 느려지며, 결과적으로 생성된 이미지의 품질이 크게 저하된다.[3]

DPDM(Differentially Private Diffusion Models)은 확산 모델 학습에 DP-SGD 를 직접 적용한 대표적인 모델 단계 방법론이다.[5] 고품질의 이미지를 생성하기 위해서는 강력한 표현력을 가진 대규모 모델이 필요하다. 그러나 DP-SGD 에서는 모델 파라미터 수에 비례하여 추가되는 노이즈가 커지게 되고, 이는 모델의 표현력을 발휘하는 데에 방해가 된다. 따라서 결과적으로 차등 프라이버시를 보장하는 모델은 표현력과 프라이버시 사이에 트레이드오프(trade-off)가 존재하게 된다. 이는 모델 단계 접근법이 갖는 내재적 딜레마로, DPDM 은 이러한 한계를 완화하기 위해 *noise multiplicity* 와 같은 기법을 제안하였다. 그러나 근본적으로 초고차원 공간에서 노이즈를 주입하는 방식의 비효율성을 극복하기는 어려움이 있다.



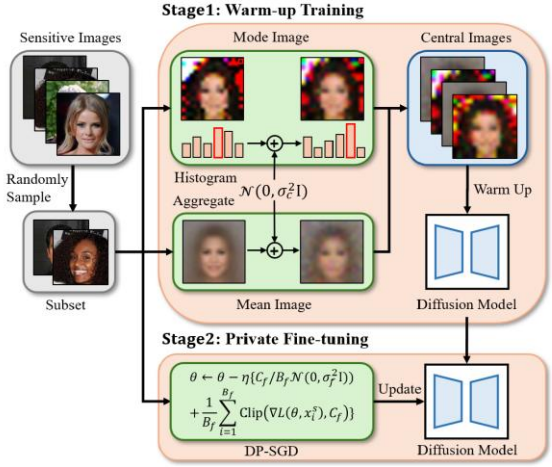
(그림 1) DPDM 의 개요.[5]

2.2. 입력 단계 프라이버시 – 데이터 공간

모델 파라미터의 초고차원 문제를 피하기 위해, 노이즈 주입의 위치를 모델 학습 이전 단계, 즉 데이터가 모델에 입력되기 전으로 옮기는 연구가 이루어지고 있다. 이 패러다임은 원본 데이터 공간 자체나 가공된 특징 공간에서 노이즈를 주입하는 방식이다. 예를 들어, 32×32 픽셀의 컬러 이미지의 CIFAR-10 데이터셋의 경우, 노이즈는 $32 \times 32 \times 3$, 즉 3,072차원의 픽셀 벡터 공간에 직접 추가될 수 있고, 이는 모델 파라미터 공간과 비교하여 훨씬 낮은 차원이다.

그 방식으로 DP-MERF(Mean Embeddings with Random Features)가 제안되었다.[6] DP-MERF 는 원본 데이터를 랜덤 푸리에 특성 공간으로 변환하고, 그 평균 임베딩에 가우시안 노이즈를 한 번만 추가하여 차등 프라이버시를 확보한 뒤, 노이즈 처리된 임베딩과 생성된 데이터 임베딩 간의 거리를 최소화해 합성 데이터를 생성한다. 또 다른 방법인 DP-NTK(Neural Tangent Kernel)는 신경망의 NTK 기반 특징을 임베딩으로 사용하여, 외부 공개 데이터나 사전학습 없이도 유의미한 표현을 얻고, 이 임베딩의 평균값에 노이즈를 추가하여 차등 프라이버시를 보장한다.[7]

DP-FETA(DP-From Easy To hArd)는 입력 단계와 모델 단계 차등 프라이버시를 결합한 접근법을 보여준다.[8] 이 방법론은 커리큘럼 학습(curriculum learning)에서 영감을 받아, "쉬운" 문제에서 "어려운" 문제로의 점진적으로 학습하는 두 단계 전략을 사용한다. 쉬운 단계에서는 민감 데이터의 부분집합으로부터 평균이나 최빈값과 같은 중심 경향 측도를 계산하여 중심 이미지(central image)를 생성하고, 가우시안 노이즈를 추가한다. 이 중심 이미지는 원본 데이터와 동일한 픽셀 차원에 존재하고, 모델은 이 데이터를 학습하여 최소한의 프라이버시 비용으로 데이터의 기본 특징을 파악한다. 어려운 단계에서는 민감한 원본 데이터와 DP-SGD 를 사용하여 fine-tuning 을 진행한다.

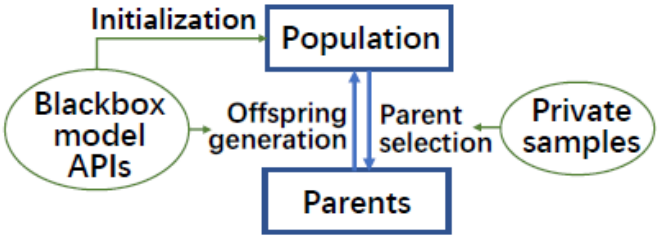


(그림 2) DP-FETA 의 개요.[8]

2.3. 출력 단계 프라이버시 – 저차원 공간

최근 AI 분야는 거대 파운데이션 모델(foundation model)의 등장으로, 모델 내부 구조에 접근하지 않고 API(Application Programming Interface)를 통해 상호작용하는 블랙박스(black-box) 패러다임으로 전환되고 있다. 이 패러다임으로 차등 프라이버시 기반의 이미지 생성 기술에서 새로운 모델을 훈련하는 대신, 이미 강력한 성능을 갖춘 기존 생성 모델을 외부에서 유도하는 방식을 사용하기 시작하였다.

PE(Private Evolution)은 블랙박스와 출력 단계 패러다임을 대표하는 알고리즘이다.[9] PE 는 진화 알고리즘의 원리를 착용하여, API 호출만으로 이미지를 생성한다. 초기에는 파운데이션 모델 API 를 호출하여 후보 이미지 집합을 무작위로 생성하고, 차등 프라이버시를 적용하여 후보 이미지 집합에서 민감 이미지와 가장 유사한 이미지를 찾는다. 이를 통해 각 후보 이미지의 최근접 이웃(Nearest Neighbor) 히스토그램을 만들고, 가우시안 노이즈를 추가하여 선택된 후보 이미지를 이용해 유사한 변형 이미지를 생성한다. 이때, 히스토그램의 차원은 후보 이미지의 개수에 불과하며, 단일 민감 데이터는 히스토그램의 한 값만 1 만큼 바꿀 수 있으므로 차등 프라이버시 민감도가 1 로 매우 낮다. 따라서, 프라이버시 예산을 가장 효율적인 저차원 공간인 의사결정 과정에만 집중적으로 사용할 수 있게 된다.



(그림 3) PE 의 개요.[9]

3. 결론 및 향후 연구 과제

이러한 차원적 시점에서의 분석은 기술 발전의 방향성을 명확하게 보여준다. 초기에 연구된 모델 단계 접근법은 초고차원 공간에 직접 노이즈를 주입하여 낮은 유용성이라는 그 한계가 명확하였다. 입력 단계 접근법은 문제의 공간을 상대적으로 낮은 데이터 특성 공간으로 옮길 수 있었고, 유의미한 개선을 이끌어냈다. 출력 단계 접근법은 파운데이션 모델을 사용하여 문제 자체를 저차원의 의사결정 공간으로 재정의함으로써 뛰어난 프라이버시와 유용성, 그리고 성능 간의 균형을 달성할 수 있었다.

이러한 차원적 관점으로 차등 프라이버시 기반의 이미지 생성 기술 연구에 중요한 시사점을 제공한다. 먼저, PE 와 같은 저차원 접근법은 파운데이션 모델의 발전에 따라, 고해상도 이미지 생성에 있어 해결책을 제시할 수 있다. 특정 데이터 유형에 더 적합하고 의미론적으로 풍부한 저차원 공간을 탐색함으로써 더 효율적인 데이터 생성 기술이 발달할 수도 있다. 또한, DP-FETA 와 같이 각 패러다임의 장점을 결합하는 하이브리드 모델에 대한 연구 방향도 고려될 수 있다.[10]

ACKNOWLEDGEMENT

이 논문은 2025 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음. 이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2023-00277326). 본 연구는 반도체 공동연구소 지원의 결과물임을 밝힙니다. 이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사드립니다.

참고문헌

[1] Samuel Yeom et al. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting", 31st IEEE Computer Security Foundations Symposium (CSF), Oxford, UK, 2018, pp.268-282.

[2] Lae Demulius et al. "Recent Advances of Differential Privacy in Centralized Deep Learning: A Systematic Survey", ACM Computing Surveys, vol.57, no.6, pp.1-28, 2025.

[3] Yinchen Shen et al. "Towards Understanding the Impact of Model Size on Differential Private Classification", arXiv, abs/2111.13895, 2021.

[4] Martin Abadi et al. "Deep Learning with Differential Privacy", Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, Vienna, Austria, 2016, pp.308-318.

[5] Tim Dockhorn et al. "Differentially Private Diffusion Models", arXiv, abs/2210.09929, 2022.

[6] Frederik Harder et al. "DP-MERF: Differentially Private Mean Embeddings with Random Features for Practical Privacy-preserving Data Generation", Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, 2021, pp.1819-1827.

[7] Yilin Yang et al. "Differentially Private Neural Tangent Kernels (DP-NTK) for Privacy-Preserving Data Generation", Journal of Artificial Intelligence Research, vol.81, pp.683-700, 2024.

[8] Kecen Li et al. "From Easy to Hard: Building a Shortcut for Differentially Private Image Synthesis", 2025 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, 2025, pp.3988-4006.

[9] Zinan Lin, et al. "Differentially Private Synthetic Data via Foundation Model APIs 1: Images", International Conference on Learning Representations (ICLR), 2024.
[10] Chen Gong et al. "DPImageBench: A Unified Benchmark for Differentially Private Image Synthesis", Proceedings of the 2025 ACM Conference on Computer and Communications Security (CCS), Taipei, Taiwan, 2025.