라벨 차등 프라이버시를 적용한 멀티모달 대조 학습 연구

김영서[1], 유민서[1], 배호[2]
[1]이화여자대학교 인공지능융합전공 석박사통합과정
[2]이화여자대학교 사이버보안전공 교수
yskim0411@ewha.ac.kr, yminsir@ewha.ac.kr, hobae@ewha.ac.kr

Multi-modal Contrastive Learning with Label Differential Privacy

Young-Seo Kim[1], Min-Seo Yu[1], Ho Bae[2]
[1]Dept. of Artificial Intelligence and Software, Ewha Womans University
[2]Dept. of Cyber Security, Ewha Womans University

요 약

최근 멀티모달 딥러닝은 모델의 높은 정확도를 보장하면서도 풍부한 지식을 학습할 수 있어 큰 관심을 받고 있다. 특히, 대조 학습을 기반으로 한 연구들이 멀티모달 딥러닝의 성능을 개선하고 있다. 그러나 멀티모달 딥러닝에서 사용하는 다중 데이터 소스가 서로 결합하여 민감한 정보를 추론하는 데 활용될 수 있으므로, 모델의 학습과정에서 프라이버시 침해의 위험성이 단일 모달 딥러닝에서보다 증가한다. 이러한 위험성은 기존 단일 모달 딥러닝에서의 프라이버시 보호 기법으로는 효과적으로 다뤄질 수 없는 까닭에 중요도가 높다. 본 논문에서는 이러한 문제를 해결하기 위해 멀티모달 대조 학습의 데이터 준비 과정에서 double randomized response 알고리즘을 활용해 라벨 차등 프라이버시를 보장하였다. 이를 통해 이미지-테이블 매칭 및 분류와 같은 주요 멀티모달 작업에서 프라이버시를 보호하면서도 80.14의 정확도를 기록하였다. 이는 데이터 보안을 고려한 멀티모달 딥러닝 모델의 성능을 처음으로 실증한 연구라는 점에서 의의가 있다.

1. 서론

최근 한 가지 데이터 형식이 아닌 이미지와 텍스트, 오디오, 테이블 등 다양한 모달리티로 모델을 학습하는 멀티모달 딥러닝이 복잡한 문제 해결에 더욱 효과적일 것이라는 기대가 커지고 있다[1]. 예를 들어, 특정 질병을 진단하는 모델을 개발하고자 할 때 이미지 데이터만을 학습하는 것보다 혈액 검사 결과 등의 테이블 데이터를 함께 활용할 경우 모델이 더욱 풍부한 지식을 학습할 것이며 모델의 정확도 또한 높아질 것이다[2]. 이러한 멀티모달 딥러닝의 장점을 살리기 위해 많은 연구[3][4][5]가 진행되었으며 그중에서도 대조 학습(Contrastive Learning, CL)을 기반으로 한 모델들의 뛰어난 성능을 보이는 것으로 알려져 있다. 그러나 멀티모달 딥러닝에서는 각 모달리티 단독으로는 드러나지 않는 개인정보가 여러 모달리티를 결합함으로써 의도치 않게 노출될 수 있다[6]. 예를 들어, 음성, 영상, 위치 정보 등을 동시에 수집하는 시스템에서는 이러한 데이터가 결합되어 행동 패턴, 선호도 나아가 건강 상태와 같은 프로필이 생성될 위험이 있다. 따라서 멀티모달 딥러닝에서는 성능뿐만 아니라 보안도 함께 고려되어야 하지만, 아직까지 멀티모달 CL에 데이터 보안을 적용한 연구는 찾아보기 어렵다.

한편 기존 단일 모달 딥러닝에서 프라이버시 문제를 해결하기 위해 차등 프라이버시(Differential Privacy, DP)[7]를 활용한 시도들이 많이 이루어져 왔다[8]. DP는 적당한 노이즈를 추가하여 통계적인 기법을 적용하여 유의미한 결과를 얻을 수 있도록 도와주며 추론 공격과 오염 공격 위험을 모두 방어할 수 있는 방법론으로 연구되고 있다[9][10]. 그러나 모달리티 간의 상관성을 적절히 고려하지 않고 단일 모달 딥러닝에서 사용된 DP를 멀티모달 딥러닝에 그대로 적용한다면, 프라이버시 위험을 효과적으로 해결하기 어려울 수 있다[11].

본 논문에서는 멀티모달 딥러닝 모델에서의 프라이버시 문제를 해결하기 위해 이미지와 테이블 데이터를 활용한 멀티모달 CL에 DP의 완화된 버전인 라벨 차등 프라이버시(label-DP)를 적용한 모델을 제안한다. 이는 멀티모달 딥러닝 모델에 label-DP를

적용한 최초의 연구 사례이다.

2. 배경

2.1 대조 학습(Contrastive Learning, CL)

CL은 자기 지도 학습의 한 방식으로, 유사한 데이터의 양성 쌍은 가깝게, 서로 다른 데이터인 음성 쌍은 멀리 위치시키는 잠재 공간을 학습한 후, 이를 기반으로 후속 작업을 수행하는 기법이다. CLIP[3]에서는 이미지와 텍스트 쌍으로 구성된 데이터셋에서 이미지 인코더와 텍스트 인코더를 통해 얻은 임베딩 값이 같은 인덱스를 가지면 양성 쌍으로, 다른 인덱스를 가지면 음성 쌍으로 설정한다. CLIP의 손실 함수는 $L = -\mathrm{SIM}(\mathrm{I}_{\text{양성}}, \mathrm{T}_{\text{양성}}) + \mathrm{SIM}(\mathrm{I}_{\text{음성}}, \mathrm{T}_{\text{음성}})$와 같이 간략하게 표현이 가능한데, 이미지와 텍스트의 양성 쌍들의 코사인 유사도는 크고, 음성 쌍들의 코사인 유사도는 작아지는 방향으로 잠재공간을 학습한 후에 새로운 이미지 데이터가 들어오면 이미지 인코더로 임베딩을 계산하여 가장 가까운 텍스트를 찾는 방식으로 활용된다.

최근에는 다양한 CL 방법들이 제안되었다. [12]는 라벨 없이 이미지의 의미 있는 표현을 학습하는 방법을 제안했으며, 클러스터 할당을 통해 시각적 특징 학습에 효과적임을 보였다. 또한, [13]은 자기 지도 학습을 비전 트랜스포머에 적용하여 기존 학습 모델에 근접한 성능을 달성했다. 멀티모달 학습 분야에서는 [14]가 비디오의 여러 모달리티 간 대조 학습을 통해 강력한 비디오 표현 학습 방법을 제시하며, 멀티모달 대조 학습의 가능성을 새롭게 보여줬다.

2.2 라벨 Renyi 차등 프라이버시(Label Renyi Differential Privacy, label-RDP)

본 논문에서는 [9]에서 소개된 Renyi 차등 프라이버시(Renyi-DP)를 적용한다. $\alpha > 1$이고 $\epsilon \geq 0$일 때, 메커니즘 $M: X \times Y \mapsto X \times Y$은 모든 $x, x' \in X$, $y, y' \in Y$에서 $D_\alpha(M(x,y) \| M(x',y')) \leq \epsilon$를 만족하면 "$(\alpha,\epsilon)$-Renyi differential privacy(RDP)를 만족한다."라고 말한다. RDP는 기존 [7]에서 소개된 DP의 확장된 버전으로, 더 정교하고 유연한 프라이버시 분석이 가능하다.

RDP를 사용한 label-DP는 [10]의 정의를 따른다. $\alpha > 1$이고 $\epsilon \geq 0$일 때, 메커니즘 $M: X \times Y \mapsto X \times Y$은 모든 $x, \in X, y, y' \in Y$에서 $D_\alpha(M(x,y) \| M(x',y')) \leq \epsilon$를 만족하면 "$(\alpha,\epsilon)$-label Renyi differential privacy(label-RDP)를 만족한다."라고 말한다. 반면, 기존 [15]에서 사용되는 label-DP는 라벨 추론 공격에 취약함을 [10]을 통해 확인되었다. 입력 데이터의 특징이 공개될 경우 라벨이 보호되지 않기 때문에, [10]에서 사용된 정의는 특징이 비공개된 라벨에 의존하는 분포에서 무작위로 추출되어 라벨 추론 공격을 방지한다.

Algorithm 1. Model-based double-RR

Given: Noise level $\lambda \in [0,1]$; feature distribution $P_{X|y}$ for each label $y \in Y$.

Input: Labeled example (x, y).

Output: Noisy labeled example $(\tilde{x}, \tilde{y})$.

1 **if** with probability $1-\lambda$ then
2 Let $\tilde{y} = y$.
3 **else**
4 Draw $\tilde{y}$ uniformly at random from Y.
5 **end if**
6 **if** with probability $1-\lambda$ then
7 Let $\tilde{x} = x$.
8 **else**
9 Draw y' uniformly at random from Y.
10 Draw $\tilde{x}$ using knn(k-nearest neighbor) algorithm from distribution $P_{X|y'}$.
11 **end if**
12 **return** $(\tilde{x}, \tilde{y})$

3. 방법

본 논문에서는 프라이버시 문제를 해결하기 위해 멀티모달 CL에 label-RDP를 적용해 보고자 한다. 이를 이미지 데이터와 테이블 데이터에 적용한다고 가정하자. 두 데이터가 페어링되기 전에 이미지 데이터에 대해 label-DP를 적용한다. 기존 연구 [10]에서 사용된 double randomized response(double-RR) 알고리즘(Algorithm 1)은 [10]의 Theorem 7.1에 의해 label-DP를 만족함을 증명되었다. 여기서 x는 이미지 데이터, y는 라벨이라고 가정한다. Algorithm 1의 1-5번째 줄은 y에 대하여 randomized response(RR)을 적용하는 부분으로서, $\tilde{y}$는 $1-\lambda$의 확률로 원래의 y가 되고, λ의 확률로 Y에서 무작위로 다른 값이 선택된다. Algorithm 1의 6-11번째 줄에서 x에 대하여 RR를 적용하는 것 또한 기본적으로 같은 방식으로 동작하되 λ의 확률로 $\tilde{x}$를 뽑을 때 자신이 선택될 확률을 높이기 위해 $P_{X|y'}$의 분포에서 균일한 확률로 뽑지 않고 거리가 k-최근접 이

웃(k-NN) 알고리즘을 사용하며 가까운 데이터 중에서 선택하도록 하였다. Algorithm 1은, $\alpha > 1$이고 P가 $X \times Y$에 대한 분포일 때, $\epsilon \geq 2\log\left(1+\frac{(1-\lambda)k}{\lambda}\right)$ 조건이 만족하면 모든 label $y \in Y$에 대해 (α, ϵ, P)-label RDP를 만족한다. 이는 [10]의 appendix E.에서 증명할 수 있다.

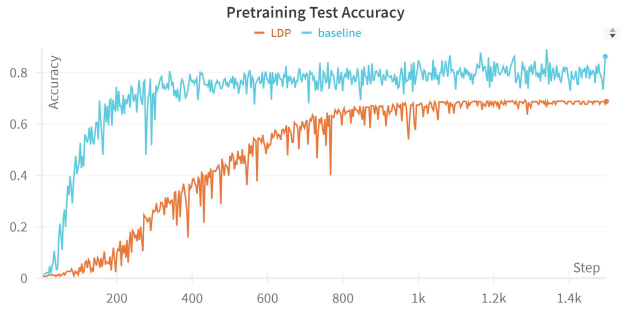
Double-RR을 통해 label-DP를 만족하는 이미지 데이터셋을 얻은 후, 동일한 인덱스를 가진 테이블 데이터셋 역시 인코더를 통해 잠재 공간에서 페어링한 후, CLIP 손실을 이용한 대조 학습을 수행한다. 테이블 데이터는 이미지 인코더를 잘 학습시키기 위한 사전 학습 단계에서만 활용되며, 최종 목표는 이미지 분류이므로 label-DP는 이미지 데이터셋에만 적용된다는 점을 확인할 수 있다.

학습은 Algorithm 1을 적용한 데이터를 기반으로 [5]의 방식에 따라 인코더를 학습하는 사전 학습 단계와 학습된 인코더를 활용한 데이터 분류를 진행하는 후속 작업 단계로 구성된다. SimCLR[4]에서는 인코더의 결과값을 입력으로 받아 두 개의 Fully Connected layer와 ReLU로 연산을 수행하는 projection head의 출력 공간에서 손실 함수를 정의한다. 따라서, 양성 쌍에 대해 각각의 projection head의 결과값 z_i, z_j가 있을 때, 손실 함수는 $l_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$로 정의되며, 이때 $\mathrm{sim}(u, v) = u^{\mathrm{T}} v / \|u\| \|v\|$로 임베딩 간의 유사도를 나타낸다.

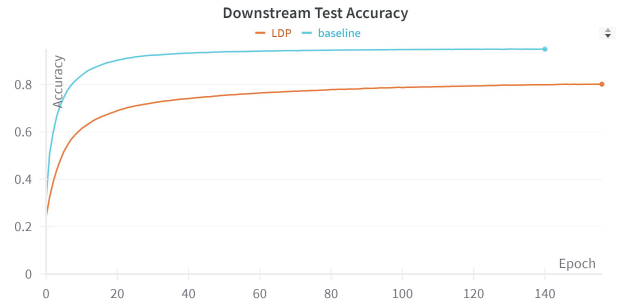
[5]에서는 CLIP[3]처럼 이미지 데이터와 테이블 데이터의 한 쌍의 양성 쌍으로 두고, 나머지 조합을 음성 쌍으로 설정하며, 손실 함수는 SimCLR[4] 방식을 따라 계산하였다.

4. 실험

실험에 사용한 데이터셋은 DVM-car 데이터셋[16]로 300x300 크기의 자동차 이미지로 구성되어 있다. 총 1,451,754개의 데이터 중 176,414개를 선정하여, 절반은 테스트 데이터로 사용하고, 나머지 절반 중 10%를 검증 데이터로, 나머지를 학습 데이터로 활용하였다. 테이블 데이터는 각 이미지 데이터와 일대일로 대응되는 데이터를 사용하였다. 실험 환경은 [5]와 동일하게 설정되었으며, 이미지 데이터는 수평 뒤집기, 45도 회전, 그리고 128x128 크롭을



(그림 1) 사전 학습 테스트 정확도



(그림 2) 후속 작업 테스트 정확도

<표 1> 멀티모달 CL 정확도

	LDP(Ours)	Non-LDP	Baseline
사전 학습	67.77	79.18	-
후속 작업	80.14	94.08	93.00

각각 0.5의 확률로 무작위로 적용하고, 밝기와 대비를 최대 0.5까지 변경하는 방식으로 데이터 증강을 수행하였다. 테이블 데이터는 너비, 길이, 높이, 휠베이스와 같은 클래스를 식별할 수 있는 열을 최대 50mm 범위에서 무작위로 변형하였고, 결측값은 반복적 다변량 대치법을 사용하여 처리하였다.

학습률은 3.00E-03, 가중치 감쇠는 1.50E-06으로 설정하였으며, 배치 크기는 512, 옵티마이저는 Adam을 사용하였다. 매 step마다 사전 학습을 통해서 잠재 공간을 학습하였고, 그 결과를 토대로 epoch마다 후속 작업으로 분류 정확도를 측정하였다. (그림 1)과 (그림 2)는 사전 학습의 세부적인 변화 추이를 보여주기 위해 step 단위로 그래프를 그렸고, 후속 작업은 epoch 단위로 x축을 구성하였다. 한 epoch는 약 12개의 step으로 구성된다.

LDP의 epsilon은 5로 설정하였으며 LDP를 적용한 모델과 LDP를 적용하지 않은 Non-LDP 모델에서의 파라미터는 모두 동일하나, early stopping으로 인해 epoch의 수는 각각 500과 444로 차이가 발생하였다. 각 단계에서의 정확도 추이는 <표 1>, (그림 1), (그림 2)에서 확인할 수 있다. [5]에서 제안된 Multimodal Imaging 모델을 Baseline 모델로 사용

하였다.

<표 1>에서의 baseline은 기존 [5] 논문에서 제시된 정확도이다. Non-LDP와 baseline은 동일한 실험 환경에서 진행되었으나, 약 1.08 정도의 차이가 나타났다. 이는 멀티 GPU 사용 및 세부 설정의 차이에 기인한 것으로 보이며, 1.08 정도의 차이만 발생한 것은 실험 세팅이 적절하게 이루어졌음을 시사한다. 최종적으로, <표 1>에서 확인할 수 있듯이 LDP를 적용한 모델과 적용하지 않은 모델 간의 정확도는 각 단계에서 10.41 및 14.84의 성능 손실이 발생하였다. 일반적으로 차등 프라이버시를 적용하면 기존 모델에 비해 어느 정도의 성능 손실이 불가피하다. 그럼에도 불구하고, 라벨 차등 프라이버시를 적용한 모델이 80% 이상의 정확도를 달성한 것은 매우 긍정적인 결과라 할 수 있다. 이는 프라이버시를 보호를 유지하면서도 성능 저하를 최소화한 학습 방법으로, 멀티모달 대조 학습에서 프라이버시와 성능 간의 균형을 성공적으로 맞췄다는 점에서 큰 의의가 있다.

5. 결론

본 논문에서는 멀티모달 대조 학습에 double randomized response 알고리즘을 적용하여 라벨 차등 프라이버시를 만족함을 확인하였다. 그러나 본 연구에서는 라벨 차등 프라이버시를 사용하여 이미지 데이터의 라벨에만 제한적으로 프라이버시 보호를 적용할 수 있었다는 한계가 있다. 즉, 제한된 값에만 프라이버시 보호가 적용된다는 점에서 한정적인 보호를 제공한다. 멀티모달 대조 학습에서 데이터 보호를 위해 라벨 차등 프라이버시를 최초로 적용한 본 연구는 의미 있는 기여를 하지만, 향후 연구에서는 특정 값에 한정되지 않고 전체 데이터셋에 대해 프라이버시를 보호할 수 있는 일반 차등 프라이버시를 만족하는 방향으로 확장할 필요가 있다.

사사

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-02068, 인공지능 혁신 허브 연구개발)

참고문헌

[1] Jabeen, S. et al, A review on methods and applications in multimodal deep learning. TOMM, 19, 2s, pp.1-41, 2023.

[2] Stahlschmidt, S.R. et al, Multimodal deep learning for biomedical data fusion: a review, Briefings in Bioinformatics, 23, 2, p.bbab569, 2022.

[3] Radford, A. et al, Learning transferable visual models from natural language supervision, PMLR ICML, 2021, pp. 8748-8763.

[5] Hager, P. et al, Best of both worlds: Multimodal contrastive learning with tabular and imaging data, CVPR, Vancouver, 2023, pp. 23924-23935.

[6] Friedland, G et al, The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3, New York, ACM, pp. 659-704.

[7] Dwork, C., Differential privacy, ICALP, Berlin, 2006, pp. 1-12).

[9] Mironov, I, Rényi differential privacy, IEEE CSF, Santa Barbara, 2017, pp. 263-275.

[10] Busa-Fekete, R.I. et al, Label differential privacy and private training data release, PMLR ICML, Hawaii, 2023, pp. 3233-3251.

[11] Cai, C. et al, A multimodal differential privacy framework based on fusion representation learning, Connection Science, 34, 1, pp.2219-2239, 2022.

[12] Caron, M. et al, Emerging properties in self-supervised vision transformers. ICCV, 2021, Montreal, pp. 9650-9660.

[13] Caron, M. et al, Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS, 2020, pp. 9912-9924.

[14] Zolfaghari, M. et al, Crossclr: Cross-modal contrastive learning for multi-modal video representations. CVPR, 2021, pp. 1450-1459.

[15] Ghazi, B. et al, Deep learning with label differential privacy, NeurIPS, 2021, pp.27131-27145.

[16] Huang, J. et al, DVM-CAR: A large-scale automotive dataset for visual marketing research and application, IEEE Big Data, Osaka, 2022, pp. 4140-4147.