修士学位論文要約（令和２年３月）

画像に対する GAN を用いたプライバシ保護データマイニングに関する研究

樺島 八入
指導教員：篠原 歩，　学位論文指導教員：全 眞嬉

A Study on Privacy Preserving Data Mining on Image data with Generative Adversarial Networks

Yashio Kabashima
Supervisor: Ayumu Shinohara,　Research Advisor: Jinhee Chun

Publishing learning model is one of important topic in machine learning era. However there are several problems such as specifying privacy information or rejection of providing privacy information. Then, privacy preserving is significant to overcome these problems. While a lot of researchers study privacy protection, there are few studies on image dataset. This is because of the huge decline of accuracy on learning model. Therefore, we reconsider the definition of privacy protection and suggest to maintain learning accuracy on protected image dataset with Generative Adversarial Networks. Finally, we confirm the quality of learning model which we suggested from experiments.

1. はじめに

機械学習モデルの共有や公開は，転移学習への応用や自分が持つモデルとの比較など様々な実タスクが可能となる一方でデータ提供者の情報特定などの危険性を持つ．そこで，プライバシ保護データマイニングと呼ばれる研究分野では差分プライバシ [1]や秘密計算などのプライバシ保護手法が提案されている．しかしながら，画像データに対してこれらを適用しても学習性能劣化や計算時間が問題となり研究が停滞している．そこで，本研究では画像データに対するプライバシ保護の定義を見直し，敵対的生成ネットワーク(GAN)を用いた生成画像を利用して学習モデルを構築することにより画像データに対するプライバシ保護と学習モデル性能維持の両立を目指す．

2. 本研究におけるプライバシ保護の定義

既存手法である差分プライバシでは適当な確率分布から発生させたノイズを元データに加えることによりプライバシ保護を行うが画像データでは高次元のためにノイズによる情報損失が大きくなるため，画像データに対する差分プライバシの適用事例が存在しない．差分プライバシは理論的保証がある一方，そこで考える攻撃者の定義が無限大の計算時間及び無限大の計算時間を所持するという非常に強力なものである．

そこで本研究では，**プライバシ保護の定義として実際の画像が訓練データとして直接使用しなければ実用的に十分であると仮定**し，GAN を用いたプライバシ保護手法を提案する．

3. 提案手法

3.1. 提案手法の概要

本研究では，世の中へ公開するための学習モデルを公開モデルと呼ぶ．提案手法は公開モデルを学習する際の訓練データの生成方法である．訓練データ生成のためには，生成画像およびラベルの付加が必要となる．また，学習性能維持のためにクラス分類が曖昧な画像を含めたいと考える．

3.2. PPML with ACGAN

ACGAN[2] とは，ラベル付き画像を生成する GAN の一種である．ACGAN により生成したラベル付き画像を訓練データとして公開モデルを構築する手法が PPML with ACGAN である．この手法の利点は，画像を生成する際にラベルを指定するためデータセットを確実に生成することができる点である．一方で，ACGAN の本来の目的はクラスごとの実際の画像に類似した画像を生成することであるため，公開モデルの性能維持に必要だと思われるクラス分類が曖昧な画像が生成されない可能性があることが欠点である．

3.3. PPML with DCGAN and Classifier

PPML with ACGAN の欠点であるクラス分類が曖昧な画像を含めるため DCGAN[3] を用いた手法を提案する．DCGAN は ACGAN のようにクラスラベルについては考慮せず画像を生成するため，クラス分類が曖昧な画像が含まれることが期待できる．しかし，DCGAN のみでは生成した画像にラベル情報が存在しないため，Classifier を

用いてラベル付加を行う．Classifier は公開モデルと同一のモデルに実際の画像を用いて学習したモデルである．DCGAN により生成した画像を Classifier に入力したときに予測されるクラスラベルを生成画像に対応するクラスラベルとして定義する．

PPML with DCGAN and Classifier の利点は，上述のようにクラス分類が曖昧な画像の生成が期待できる点であるが，欠点としてはラベルの指定ができないために，生成画像に偏りが生じると訓練データの生成が難しくなる可能性がある点があげられる．

4. 実験

①プライバシ保護を行わない訓練データ（ベンチマーク）②PPML with ACGAN により生成する訓練データ③ PPML with DCGAN and Classifier により生成する訓練データの 3 ケースで構築した公開モデルの評価を行う．実験タスクは，MNIST を用いた手書き数字画像の 10 クラス分類，UTKFace と呼ばれる顔画像データセットを用いた性別に関する 2 クラス分類，CIFAR-10 を用いた動物に関する 4 クラス分類である．

表 1 実験結果

	ベンチマーク	PPML with ACGAN	PPML with DCGAN and Classifier
MNIST	99.3%	96.9%	98.9%
UTKFace	90.1%	74.1%	86.6%
CIFAR-10	88.6%	71.7%	76.3%

実験結果を表 1 に示す．まず，2 つの提案手法の正答率を比較すると，全ての実験で PPML with DCGAN and Classifier の正答率が PPML with ACGAN を上回ったことがわかる．また，プライバシ保護を行っていないベンチマークと比較すると，PPML with ACGAN は性能劣化が見られる一方 PPML with DCGAN and Classifier は性能劣化が抑制できている．これは，クラス分類が曖昧な画像の有無が影響していると考えられる．ACGAN ではクラス分類が曖昧な画像が生成されにくいため，訓練データに対しては学習がすぐに収束するが，決定境界の精度が低くテストデータに対して正答率が低下したのではないかと考えている．

また，CIFAR-10 を用いた実験では 2 つの提案手法で学習性能が劣化した．この理由として CIFAR-10 の画像データは，クラスラベルの対象物体のみならず，背景などの情報も含まれた画像データであるため，GAN による画像の生成精度が低下したことが原因として考えられる．

以上の考察について検証するため，図 1 で各提案手法での 3 つのデータセットに対する生成画像を示す．各行がクラスラベルに対応している．

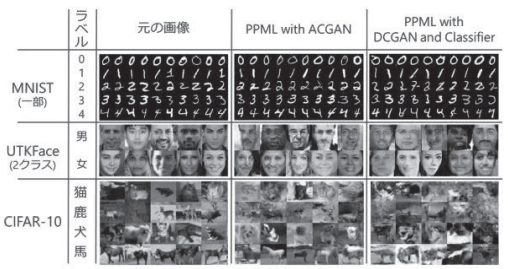

図 1．GAN による生成画像サンプル

生成画像サンプルの各行を見ると，PPML with ACGAN により生成された画像では他のクラスに属しても良いようなクラス分類が曖昧な画像は見られないが PPML with DCGAN and Classifier では，クラス分類が曖昧な画像も存在することが確認できる．また CIFAR-10 による実験ではどちらの手法においても生成精度が他のデータセットの場合よりも低下していることが確認できる．

5. 結論・今後の課題

実験結果より，PPML with DCGAN and Classifier はプライバシ保護と性能維持の観点で一定の成果を確認することができた．一方，本研究ではプライバシ保護の定義として実際の画像を公開する学習モデルの訓練データに用いないこととしたが，この定義の妥当性について検証する必要がある．また，CIFAR-10 のような複雑な画像では GAN 生成精度が劣化したため，より高性能な GAN を用いることにより生成精度の劣化を抑制することが今後の課題としてあげられる．

参考文献

1) Dwork et al. Differential Privacy, ICALP, 2006
2) Odena et al. Conditional image synthesis with auxiliary classifier gans. ICML, 2017
3) Radford et al. Unsupervised representation learning with deep convolutional generative adversarial networks. ICLR, 2016