G-7

Membership Inference Attacks に対する DP-SGD 及び Dropout の効果検証

Verification of the Effects of DP-SGD and Dropout on Membership Inference Attacks

○宮野咲紀[1]，　関弘翔[2]，　細野裕行[2]
*Saki Miyano[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: The purpose of this study is to propose an effective defense method against MIA. In this study, we examined the effectiveness of the methods that are expected to have defense capability.

1. まえがき

械学習研究には一般に大量の学習データが必要であるが，近年，そのデータの取り扱いに関する意識が高まっている。特に顔画像や医療データをはじめとする，個人に結びつきやすいデータを用いて学習を行う際には，それらのプライバシーが問題となっている。

その一例として，あるデータが学習用データセットに含まれていたかどうかを推測する「Membership Inference Attacks(MIA)」[1]の可能性が示唆されており，データを取り扱ううえで重大なリスクになる。

本研究の目的は MIA に対して有効な防衛手法を提案することであり，本報告では防衛能力が期待される手法の効果検証を検討した。

2. 関連研究と本研究の位置づけ

学習データのプライバシー保護の実現に有力とされる技術に Differential Privacy[2]がある。

Differential Privacy とは，データに Gaussian Noise などを加えて扱うことで一つ一つのデータをぼかし，その結果どの程度プライバシーが保護されているのかを定義した概念である。

MIA に対する防衛法として，モデルの最適化に用いられる Stochastic Gradient Descent（SGD）に DP の概念を取り入れた Differential Private–SGD (DP-SGD)[3]を適用する効果を検証した研究[4]がある。DP-SGD は，SGD を用いた最適化における勾配を一定のノルムで clip し，そこに Gaussian Noise を加えることで DP を保証するものである。また，別の防衛法として，モデルの過学習を抑制する機構の一つである Dropout を適用する効果を検証した研究[5]がある。

しかし，これらの防衛法を統合的に扱った際の効果検証がなされていないため，実用上問題がある。本研究では DP-SGD と Dropout を用いて，データセットへの攻撃手法の一つである MIA に対する効果の検証を行う。

3. 検証方法

本研究では，CIFAR-10 などの一般的な画像分類用データセットを対象として画像分類用 CNN モデルを構築する。この際に，何も防衛手段を施さないモデル，様々な強度で DP-SGD を適用させたモデル，ネットワークの一部をランダムに欠落させる Dropout を様々な割合で適用させたモデル，DP-SGD と Dropout を組み合わせたモデルの構築を行う。これらのモデルに対して MIA を実行し，画像分類器としての分類精度，及び MIA に対する頑健性を調査する（Fig. 1）。

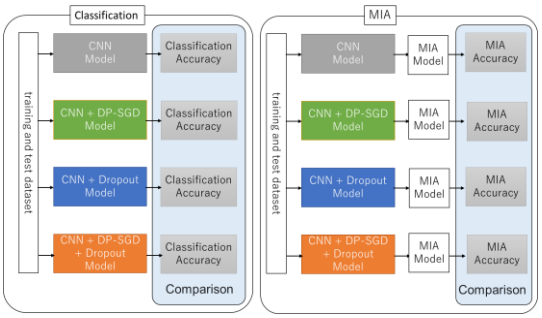

Figure 1. Schematic diagram of verification

4. まとめ

本研究では DP-SGD と Dropout を用いて，データセットへの攻撃手法の一つである MIA に対する効果の検証のためのモデル構築を行った。

参考文献

[1] R. Shokri, et al. : “Membership inference attacks against machine learning models,” 2017 IEEE Symposium on Security and Privacy (SP). IEEE, (2017)
[2] C. Dwork. : “Differential Privacy, ” In ICALP 2006, (2006)
[3] M. Abadi, et al. : “Deep learning with differential privacy,” In 23rd ACM Conference on Computer and Communications Security(ACM CCS), pp. 308-318, (2016)
[4] M. A. Rahman, et al. : “Membership Inference Attack against Differentially Private Deep Learning Model,” TRANSACTIONS ON DATA PRIVACY 11, pp.61-79, (2018)
[5] A. Salem, et al. : “ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models,” arXiv:1806.01246v2, (2018)

１：日大理工・院（前）・情報，　２：日大理工・教員・情報