Research Proposal——Add privacy protection in Generative Adversial Networks(GAN)

Haihan Gao
Xiaolu Chen
Li Zhang

October 10, 2022

Generative Model has been used to generate fake data from the distribution of real data. Some impressive work has been done with data generation such as Generative Adversarial Network(GAN) and Variation Autoencoder(VAE). An universial idea of building a generative model is to map real data(such as images and texts) into latent variables which represent the information or cluster characteristic in latent space. Then we use latent variables from encoder as the parameter of the target distribution.Sampling from the target distribution will generate new data which retains features from origin data as well as contains differences compared with origin input. In order to use Gradient descent to optimize deep model,we need to design loss function predently.Generally the loss function is made up with two parts.The first one is difference with input,such as mseloss.Another one is KL divergence between latent distribution and a priority assumption. If we give the first part a high priority,we will find that our model tends to "copy" input data into output data.This may result into leakage of training data.

1 Introduction of the problem

In this section the essence of the proposed work is described by answering several key questions.

Main questions related to our work

Machine Learning Tasks,including regression,classification and other ML task ,require a large amount of Training Data.Generally we collect data from real world.

In some specific scenarios,data is expensive and contains some privacy informations,such as human face and fingerprint. An alternative solution to this problem is to build a model which adds noise with real data.This model is called generaitve model. However,there is a tradeoff between data diversity and data utilization.Diversity means generative data should be different from real data.Suppose a extreme scenarios, if the generative model only copy input images to output images,there will be little significance to introduce generative model.Utilization means the output should be able to train downstream model in order to achieve our goals,such as image classification.

Solution to the problem

Based on the work of differential privacy in deep learning,we add noise in gradient when updating parameter of networks.GAN is consisted with Generative Net and Discrimination Net and they are trained with a max-min loss.
According to the structure of GAN,the Discrimination net is removed after the whole net is well trained. only add noise when gradient is passed from Discrimination to Generative net.This well help us set better noise range ϵ

2 Preliminary notation and background

We introduce some notations

1. DP-SGD Training GANs with the DP-SGD method can be effective in generating high-dimensional sanitized data. However, DP-SGD relies on the clipping bound of gradient norm, i.e., the sensitivity value. Sensitivity values vary greatly with model architecture and training dynamics, which makes the implementation of DP-SGD difficult.

2. PATE Privatization of Teacher Aggregates (PATE) has recently been adapted to generative models. and two main approaches were studied: PATE-GAN and G-PATE. Both of these methods only train generators with DP guarantees, however, the gradients of G-PATE need to be manually selected to adapt to the PATE framework and secondly the high-dimensional nature of the PATE framework brings high privacy costs.

3. Fed-Avg GAN Fed-Avg GAN is a solution to the decentralized case by using the DP-Fed-Avg algorithm to adapt GAN training to provide user-level DP guarantees under a trusted server. And it merely works on decentralized data.

3 Relevant Related Work

In this section, identified related work is described.

References

[1] Abadi, Martin and Chu, Andy and Goodfellow, Ian and McMahan, H Brendan and Mironov, Ilya and Talwar, Kunal and Zhang, Li,Deep learning with differential privacy,Proceedings of the 2016 ACM SIGSAC conference on computer and communications security

[2] Chen D, Orekondy T, Fritz M. Gs-wgan: A gradient-sanitized approach for learning differentially private generators[J]. Advances in Neural Information Processing Systems, 2020, 33: 12673-12684.