GR-70 Preventing data reconstruction with adaptive augmentation

Abstract

In privacy-preserving deep learning, differential privacy is widely used to protect sensitive data by injecting noise in the training process. However, many researchers struggled for balancing privacy and utility as the amount of noise significantly affects the performance. In our research, we introduce a data augmentation-based privacy-preserving strategy against the leakage of data using model gradients. Our research utilizes model accuracy and attacks accuracy as a metric for comparison, which indicates the accuracy of an augmented dataset and the accuracy of reconstructed images with augmentation applied. Compared to the differentially private stochastic gradient descent model introduced by Abadi et al., (2016), Our model shows superior accuracy in CIFAR-10 and finds adaptive augmentation per label that guarantees the best performance.

Introduction

As privacy becomes an issue in ML/DL, interests in differential privacy are significantly growing. However, differential privacy in deep learning is hard to implement and requires complicated background knowledge. As the main idea of differentially private deep learning is injecting noise in the training process to protect the sensitive data, we thought distorting images in the dataset through augmentation might give a similar effect compared to differential privacy. Our research focuses on finding the best augmentations and magnitude for each label in a dataset such as CIFAR-10 to compare the efficiency with differentially private deep learning models. To facilitate the augmentation experiment in an organized manner, we utilized the augmentation schemes introduced in RandAugment (2019).

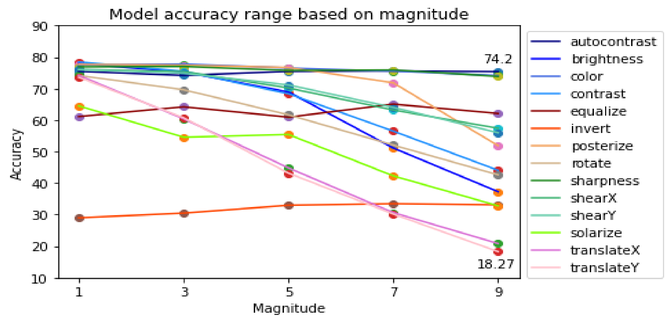
Research Question(s)

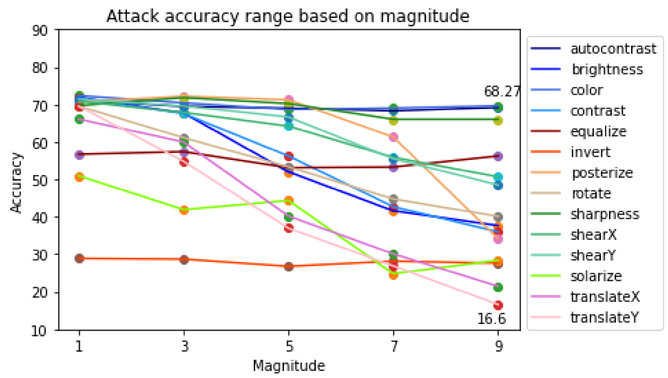
- How can sensitive information be reconstructed from the gradient of models in ML/DL?
- How is differential privacy used to protect sensitive data? What are issues regarding differential privacy?
- How can image augmentation be used to prevent the leakage of sensitive data?

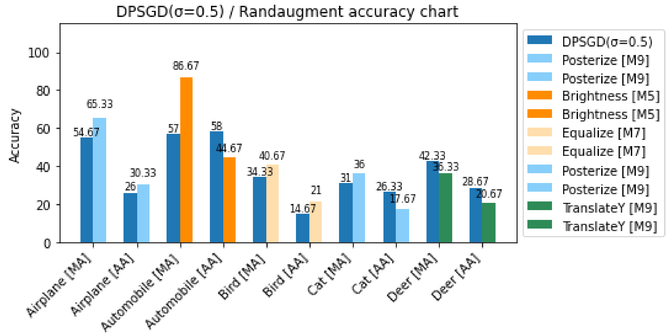
Materials and Methods

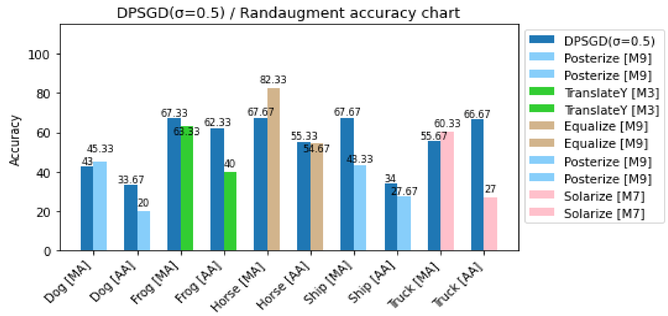
RandAugment consists of 14 different augmentation schemes and magnitude as a parameter that adjusts the level of augmentation. Instead of randomly applying augmentation schemes to the image dataset, we applied each augmentation to the dataset to find the best augmentation and magnitude per label. In our research, we implemented the DP-SGD model from Abadi et al., (2016) using Pytorch to compare the accuracy. With the DP-SGD model, we utilized ResNet-18 model with three different amounts of noise, σ=0.1, 0.5, 1.0 each, and set the result of σ=0.5 as a comparison group. From the five times of accuracy measurement, DP-SGD(σ=0.5) with resnet18 shows 52.06% model accuracy and 40.57% of attack accuracy on average. As an experimental group, we measured the model accuracy and attack accuracy of 14 different augmentations with magnitude 1,3,5,7,9. The distortion of the image gets severe as the magnitude increases. Through the comparison, we have found the best augmentation schemes that apply to 10 different CIFAR-10 labels. For example, the “automobile” label has 86.67% model accuracy and 44.65% attack accuracy when the augmentation “brightness” with magnitude 5 is applied. The comparison group shows 57.2% of model accuracy and 58.0% of attack accuracy when σ=0.5. The augmentation schemes that guarantee the lowest attack accuracy indicate that data reconstruction can be prevented by using data augmentation, and it provides better performance than DP-based defense strategies.

Results









Conclusions

We devised an idea that image augmentation may be used to protect privacy in deep learning like differential privacy. Previous approaches using differential privacy require complicated implementation and balance between privacy and utility, but our strategy can be applied directly to a dataset in the classification process in a simple manner. Our results show that giving distortions on the image through augmentation can give similar effects on protecting sensitive data in ML/DL. In future research, we will find an automatic search algorithm that optimizes the best augmentation and magnitude based on image labels.

Acknowledgments

I would like to thank my advisor Dr. Junggab Son for his advice and support, and my coworker Hongkyu Lee for tips on implementing algorithms.

Contact Information

Seunghyeon Shin – sshin9@students.Kennesaw.edu
Junggab Son – json4@kennesaw.edu

References

- M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov,K. Talwar, and L. Zhang, “Deep learning with differential privacy,” inProceedings of the 2016 ACM SIGSAC conference on computer andcommunications security, 2016, pp. 308–318.
- E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment:Practical automated data augmentation with a reduced search space,”inProceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition Workshops, 2020, pp. 702–703
- J. Geiping, H. Bauermeister, H. Dr ̈oge, and M. Moeller, “Invertinggradients–how easy is it to break privacy in federated learning?”arXivpreprint arXiv:2003.14053, 2020.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author(s) Seunghyeon Shin
Advisors(s) Junggab Son