Combining SISA Exact Machine Unlearning with Differential Privacy

UNIVERSITY OF TORONTO

Tobias Klausen, Tahmid Mostafa

Key Question

How can we combine exact machine unlearning via SISA and differential privacy (DP) without sacrificing prediction accuracy?

Introduction

Neural networks are often trained on sensitive user data which can leak. Vanilla SISA does not address privacy of existing training data.

Our contributions are:

- We limit leakage of private data using differential privacy [3]
- We utilize exact machine unlearning via the SISA [1] framework to allow the user to request deletion of their data
- We use transfer learning to mitigate the accuracy degradation introduced by DP and SISA to be able to train truly deep neural networks

We investigate the impact on accuracy wrt.:

- Different numbers of shards
- Privacy budgets
- Finetuning methods

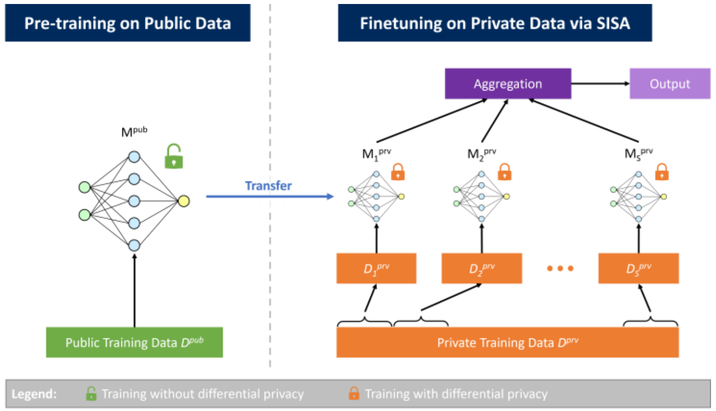
This will allow us to suggest a finetuning method yielding the best accuracy given a number of shards and a privacy budget.

Background

- **SISA:** Dataset is divided into S disjoint shards and one constituent models is trained per shard in isolation. To generate prediction results, the prediction vectors of all constituent models are averaged.
- **Differential Privacy:** Applied to SGD during finetuning with given privacy budget ε
- **Transfer Learning:** Machine learning method where a model trained for one specific task is used as the starting point for training a model on another task. This can improve accuracy in cases where training capabilities are limited.

Our Framework

Overview

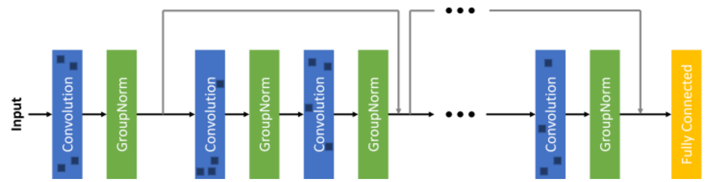


1. Pre-train one neural network M^{pub} on public data without differential privacy or SISA.
2. Copy this network S times (M_1^{prv}, …, M_S^{prv}) and use them as a starting point for the finetuning on private data within the SISA framework using differential privacy.

Note: For simplicity, we use SISA without slicing as the expected privacy and classification accuracy implications are expected to be minimal.

Finetuning

It is important to carefully select trainable parameters to improve the privacy-accuracy tradeoff according to Luo et al. [2]



We investigate 3 different finetuning methods on GroupNorm-ResNet18:

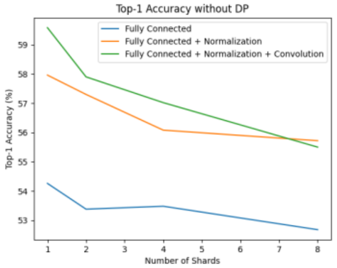
1. Finetune fully connected layer only
2. Finetune fully connected and group normalization layers
3. Finetune fully connected and group normalization layers as well as parameters of convolution layers **with large magnitude**

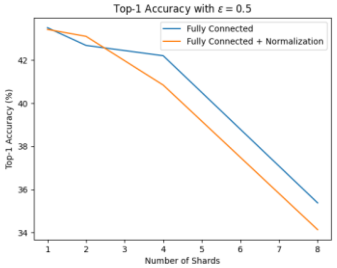
Experiments

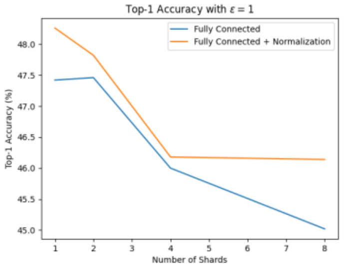
Setup

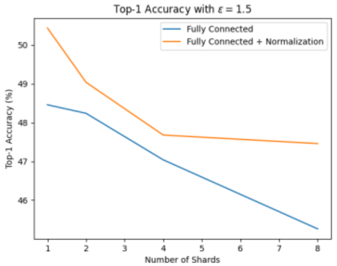
We focus on image classification tasks using ResNet18, similar to [1] and [2]. We run the following transfer learning experiment:
Pre-training: CIFAR-100 → **Finetuning:** CIFAR-10

Results

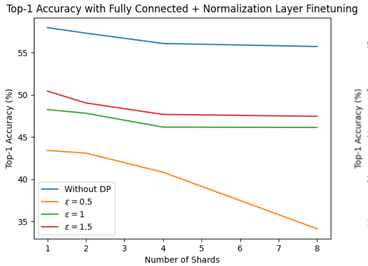


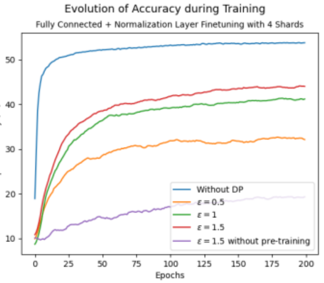






- Without DP and with moderate privacy budget (ε = 1, 1.5), the accuracy improves with more finetuning as expected
- For a very low privacy budget (ε = 0.5), more finetuning hurts accuracy as more noise is added to each finetuned parameter





- The higher the privacy budget ε, the higher is the overall accuracy
- Convergence during training is slower with DP
- Transfer learning is necessary

Conclusion

Transfer learning with carefully selected finetuning methods makes the combination of exact machine unlearning via SISA and differential privacy viable.

Limitations

- Experiments with differential privacy and partial convolution layer finetuning couldn't be executed due to hardware limitations
- Experiments with additional datasets (i.e. Pre-training: ImageNet → Finetuning: CIFAR-100) could be performed to investigate the behaviour of our framework on very little training data per target class
- Experiments with different unlearning requests were not conducted due to limited time. However, we expect the same behaviours with lesser accuracy as data-points are removed from shards.

References

[1] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2020). Machine Unlearning. ArXiv:1912.03817 [Cs]. https://arxiv.org/abs/1912.03817

[2] Luo, Z., Wu, D. J., Adeli, E., & Fei-Fei, L. (2021). Scalable Differential Privacy With Sparse Network Finetuning. Openaccess.thecvf.com. https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Scalable_Differential_Privacy_With_Sparse_Network_Finetuning_CVPR_2021_paper.html

[3] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. https://doi.org/10.1145/2976749.2978318