Differential Privacy in Image Classification using ResNet-20 and DP-SGD Optimization

Praveen Rangavajhula
Department of Computer Science
University of Georgia
Athens, GA, 30602
praveen.rangavajhula@uga.edu

Alexander Darwiche
Department of Computer Science
University of Georgia
Athens, GA, 30605
alexander.darwiche@uga.edu

Deven Allen
Department of Computer Science
University of Georgia
Athens, GA, 30605
dca09692@uga.edu

Abstract

This project proposes a differentially private image classification system using ResNet-20 and Differentially Private Stochastic Gradient Descent (DP-SGD). We explore both non-private optimizers and DP versions, justifying the choices based on prior work and their potential to outperform others. The project will focus on achieving competitive accuracy while satisfying privacy guarantees.

1 Introduction

The proliferation of machine learning models, especially in image classification, has raised concerns over privacy. Differential privacy (DP) addresses these concerns by ensuring that the model does not inadvertently leak sensitive information about individual data points. In this proposal, we focus on the ResNet-20 model, which is suited for tasks like CIFAR-10, and aim to implement and improve DP-SGD, an existing algorithm for training deep learning models with privacy guarantees.

2 Motivation and Problem Statement

In many real-world applications, such as healthcare and finance, privacy concerns hinder the deployment of machine learning models. Current state-of-the-art models like ResNet-20 achieve high accuracy but are vulnerable to attacks that could leak sensitive information. By incorporating differential privacy through the use of DP-SGD, we aim to develop a model that balances both accuracy and privacy.

3 Methodology

3.1 Model Architecture: ResNet-20

We will implement the ResNet-20 model for CIFAR-10, a standard dataset for image classification tasks. Although ResNet-18 is more common, we chose ResNet-20 for its deeper structure, making it a suitable candidate for our privacy-preserving project. If necessary, we may modify the architecture slightly to optimize for DP compatibility.

Preprint. Under review.

3.2 Non-private Optimizers to Try

We will first experiment with non-private optimizers to establish baseline performance for ResNet-20 on CIFAR-10. The following optimizers will be explored:

- **SGD:** Standard Stochastic Gradient Descent for baseline comparison.
- **Adam:** Adaptive Moment Estimation for better convergence in non-private settings.
- **RMSprop:** To explore gradient normalization impact in non-private training.

3.3 Differentially Private Optimizer: DP-SGD

We will implement DP-SGD as our privacy-preserving algorithm. The key components of DP-SGD are:

- **Gradient Clipping:** Limits the influence of individual examples during training.
- **Noise Addition:** Adds noise to gradients to ensure privacy (via Opacus or TensorFlow Privacy libraries).
- **Privacy Accounting:** We will use Rényi Differential Privacy (RDP) for privacy budget tracking.

3.4 Incremental Improvements

We plan to explore the following enhancements:

- Alternative noise mechanisms and their effect on utility-privacy tradeoffs.
- Adaptive gradient clipping methods that dynamically adjust the clipping threshold.
- Experimenting with different optimizers (e.g., DP-Adam).

3.5 Rationale for Choosing DP-SGD

- DP-SGD provides well-documented privacy guarantees (Abadi et al. 2016) while maintaining decent utility for image classification tasks.
- The addition of noise and gradient clipping help ensure (ϵ, δ)-differential privacy, making it ideal for sensitive applications.
- Previous work shows that DP-SGD, when optimized, can yield near state-of-the-art accuracy for differentially private models (De et al. 2022).

3.6 Why This Approach Will Outperform Others

- Our approach leverages the simplicity of ResNet-20, optimized for smaller datasets like CIFAR-10, combined with DP-SGD, a proven differential privacy technique.
- By experimenting with different gradient clipping techniques and noise scales, we aim to find an optimal trade-off between accuracy and privacy.
- We will investigate modifications like adaptive clipping to enhance performance.

3.7 Pseudo-code for Non-Private SGD

Below is a simplified pseudo-code for the non-private SGD we plan to privatize:

```
for each batch (X, y):
    pred = model(X)
    loss = loss_fn(pred, y)
    loss.backward()
    optimizer.step()
    optimizer.zero_grad()
```

4 Experimental Setup

4.1 System Description

We will use PyTorch for model implementation and training. The DP-SGD implementation will be based on the Opacus library. Training will be performed on GPUs available via our departmental server.

4.2 Dataset

We will use the CIFAR-10 dataset, consisting of 60,000 32x32 RGB images, which is commonly used for image classification tasks. The dataset is built-in in PyTorch, and we will load it using standard libraries.

4.3 Metrics

- **Training Loss/Accuracy:** Standard accuracy on CIFAR-10 for both non-private and private models.
- **Privacy Budget:** We will measure (ϵ, δ) using RDP to ensure privacy compliance.
- **Efficiency:** Time complexity and memory usage will be tracked.

4.4 Baseline Models

We will compare the performance of DP-SGD with non-private models and other private optimizers like DiceSGD test.

5 Related Work

Several approaches to differentially private deep learning have been explored in the literature. **?** introduced DP-SGD, which remains the most widely used technique for privacy-preserving model training. Recent works like **?** have explored scaling DP training for improved accuracy, which is also one of our goals.

6 Timeline and Milestones

- Week 1-2: Implement ResNet-20 for CIFAR-10.
- Week 3-4: Integrate DP-SGD using the Opacus library and test on CIFAR-10.
- Week 5-6: Experiment with incremental improvements and alternative optimizers.
- Week 7-8: Final evaluation and report preparation.

7 Conclusion

Through this proposal, we aim to develop a differentially private deep learning model that can achieve state-of-the-art accuracy on CIFAR-10 while providing strong privacy guarantees. We will explore DP-SGD as a baseline and investigate potential improvements in both privacy and efficiency.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318).

[2] De, S., Berrada, L., Hayes, J., Smith, S. L., & Balle, B. (2022). Unlocking high-accuracy differentially private image classification through scale. arXiv preprint arXiv:2204.13650.

GitHub link sharing?