Differentially Private Stochastic Gradient Descent

Julian Matthews John Foster Jody Sunray

CSCI-4968: ML and Optimization

April 24, 2023

Presentation Overview

Motivation

Many of the most powerful machine learning models require access to large and representative datasets that may contain sensitive data, such as patient records.

- Deep learning models are often vulnerable to privacy attacks.
- Complex models also have a tendency to ”memorize” data.

How can we make sure these models do not expose private information?

The solution is differential privacy!

Differential Privacy

Differential privacy (DP) ensures that no single data point significantly impacts a learned model.

- In other words, the output of a model trained with and without a particular data point should be the same.
- The idea is that we add just enough noise such that the privacy of the data is preserved.

Stochastic Gradient Descent

Stochastic gradient descent (SGD) is a variation of gradient descent where a parameter update is performed for each training point x_i.

Definition 1.1 (Stochastic Gradient Descent).

The update step for stochastic gradient descent is defined as follows:

$$\theta_{i+1} = \theta_i - \eta \nabla_\theta \mathcal{L}(\theta_i; x^{(i)}, y^{(i)}),$$

where $\mathcal{L}$ is the model objective, $x^{(i)}$ is a training data point, $y^{(i)}$ is a training label, and η is the stepsize we will take in each update.

Pure Differential Privacy

Definition 1.2 (Pure differential privacy).

Let D and D' be two datasets that differ by one element. We say that a randomized algorithm $\mathcal{M}$ is ε-differentially private if the probability that a particular set of outcomes S is observed never differs by more than $\exp(\varepsilon)$ between D and D'. In other words,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in S].$$

Approximate Differential Privacy

We often relax this definition of differential privacy to include a failure probability term, which is significantly smaller than the probability of any given outcome.

Definition 1.3 (Approximate differential privacy).

Let δ be a failure probability for the differential privacy definition, such that with probability δ, we get no privacy guarantee. Extending the original definition we have

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in S] + \delta,$$

where δ is very small. We say that $\mathcal{M}$ is (ε, δ)-differentially private.

Differentially Private Stochastic Gradient Descent

Differentially private stochastic gradient descent (DP-SGD) is an extension of SGD that incorporates differential privacy.

- Provided the definition of DP, let $\mathcal{M}$ be the learning algorithm, D be the training dataset, and $\mathcal{M}(D)$ be the resulting ML model.
- The intuition behind DP-SGD is it ensures that model parameters do not have too much information about the data.
- To achieve this, we clip the gradients and add noise to each gradient update. This places a bound on the amount of privacy lost per gradient update.

Sensitivity

The sensitivity of a function determines the amount of noise that needs to be added to ensure it is ε-differentially private.

- In DP-SGD, we add noise to the updates based on the sensitivity of the gradients.

Broadly speaking, sensitivity is the maximum amount by which the output of a function can change when the input data is changed by one entry, such as by removing a single data point.

Global Sensitivity

Global sensitivity considers the maximum difference between any two neighboring datasets. Thus, this measure of sensitivity is independent of the particular dataset we are querying.

Definition 2.1 (Global sensitivity).

Given a query function f, the global sensitivity of f is the maximum difference in output, considering all possible datasets that differ by one entry. In other words,

$$\Delta f_{\text{GS}} = \max_{D_1, D_2} ||f(D_1) - f(D_2)||,$$

where D_1 and D_2 are any two datasets that differ by one element.

Local Sensitivity

On the other hand, local sensitivity refers to the maximum distance between two neighboring datasets when one of them is fixed.

Definition 2.2 (Local sensitivity).

Given a query function f and a known dataset D_1, the local sensitivity of f is the maximum difference in output if D_1 is changed by one entry. In other words,

$$\Delta f_{\text{LS}} = \max_{D_2} ||f(D_1) - f(D_2)||,$$

where D_1 is some known dataset and D_2 is a dataset that differs from D_1 by one element.

Laplace Mechanism

The Laplace mechanism adds just enough random noise, sampled from the Laplace distribution, to satisfy pure differential privacy. It is commonly used for real-valued functions $f : D \to \mathbb{R}$.

Definition 2.3 (Laplace mechanism).

Let $f(D)$ be some query function that takes D as the input dataset. To make the function ε-differentially private, we add Laplace noise proportional to the sensitivity of f and inversely proportional to the privacy loss ε. In other words, the following definition of $F(D)$ is ε-differentially private:

$$F(D) = f(D) + \text{Lap}\left(\frac{s}{\varepsilon}\right),$$

where s is the sensitivity of f.

Gaussian Mechanism

On the other hand, the Gaussian mechanism only satisfies approximate differential privacy, and it is preferred for vector-valued functions $f : D \to \mathbb{R}^k$.

Definition 2.4 (Gaussian mechanism).

Let $f(D)$ be some query function that takes D as the input dataset. To make the function (ε, δ)-differentially private, we add Gaussian noise with variance $\sigma^2 s^2$, where $\sigma^2 = \dfrac{2\log(1.25/\delta)}{\varepsilon^2}$ and s is the sensitivity of f. In other words, the following definition of $F(D)$ is (ε, δ)-differentially private:

$$F(D) = f(D) + \mathcal{N}(\sigma^2 s^2).$$

The Algorithm

Algorithm 1 DP-SGD (Abadi et al.)

Input: Examples $x_1, \ldots, x_N$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, minibatch size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take random sample L_t with probability L/N

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{||\mathbf{g}_t(x_i)||_2}{C}\right)$

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

end for

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Fundamental Law of Information Recovery

On every iteration of DP-SGD, more of the data is accessed, and as a result there is more privacy loss. This phenomenon is known as the Fundamental Law of Information Recovery.

Definition 4.1 (Fundamental Law of Information Recovery).

If we query a dataset enough times, we can discover information about specific data points. In other words, the privacy loss increases with the number of queries.

Sequential Composition

One important property of differential privacy that follows from this notion is that we can add up privacy budgets when two mechanisms sequentially access the same input data. This is called sequential composition.

Theorem 4.1 (Sequential Composition).

Let $\mathcal{M}_1$ be an $(\varepsilon_1, \delta_1)$-differentially private mechanism, and let $\mathcal{M}_2$ be an $(\varepsilon_2, \delta_2)$-differentially private mechanism. Then the combination of $\mathcal{M}_1$ and $\mathcal{M}_2$, given by $\mathcal{M}_{1,2} = (\mathcal{M}_1, \mathcal{M}_2)$, is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private.

Privacy Amplification

We can actually reduce the privacy loss of an ε-differentially private mechanism by only accessing a fraction q of a dataset. This gives rise to the privacy amplification theorem, which guarantees a privacy loss of $q\varepsilon$. Intuitively, this guarantee makes sense since we are only accessing a subset of the data.

Theorem 4.2 (Privacy amplification).

Let $\mathcal{M}$ be an (ε, δ)-differentially private mechanism. If we query a fraction q of a dataset D, the privacy loss reduces by a factor of q and we say that $\mathcal{M}$ is $(q\varepsilon, q\delta)$-differentially private.

Parallel Composition

We can combine the sequential composition and privacy amplification theorems to motivate the parallel composition theorem, which is based on splitting a dataset into disjoint chunks and then running an ε-differentially private mechanism on each chunk.

Theorem 4.3 (Parallel Composition).

Let $\mathcal{M}$ be an (ε, δ)-differentially private mechanism, and let D be a dataset that has been split into k disjoint chunks such that $D = d_1 \cup d_2 \cup \cdots \cup d_k$. Then the combination of $\mathcal{M}(d_1), \mathcal{M}(d_2), \cdots, \mathcal{M}(d_k)$ is (ε, δ)-differentially private.

Privacy Accounting

Privacy accounting refers to estimating the total privacy loss of training a differentially private model using DP-SGD.

Using sequential composition and privacy amplification, the overall total privacy budget of the algorithm can be computed as $(qT\varepsilon, qT\delta)$.

While this privacy bound is guaranteed to hold true, it is very loose.

Strong Composition

There has been a substantial amount of research dedicated to privacy loss composition, and this research has led to the strong composition theorem, which places a much tighter bound on privacy.

Theorem 4.4 (Strong composition).

Let $\mathcal{M}$ be an (ε, δ)-differentially private mechanism that takes a dataset D as input. If we query a fraction q of D over T iterations, we get a $(q\varepsilon\sqrt{T\log(\frac{1}{\delta})}, qT\delta)$-differential privacy guarantee.

Moments Accountant

We can actually get an even tighter bound given by the moments accountant privacy guarantee (Abadi et al.), which can be used in Algorithm 1 to compute an appropriate overall privacy cost.

Theorem 4.5 (Moments Accountant).

Let $\mathcal{M}$ be a (ε, δ)-differentially private mechanism that takes a dataset D as input. If we query a fraction q of D over T iterations, we get a $(q\varepsilon\sqrt{T}, \delta)$-differential privacy guarantee.

Convergence Guarantees

- For any arbitrary neural network given the loss dynamics, we can construct the neural tangent kernel (NTK) matrix.
  - The NTK matrix is a symmetric, positive semi-definite (PSD) matrix that captures information about the loss function of a neural network.

The NTK matrix is crucial to analyzing convergence behavior as breaking its positivity worsens convergence.

Local vs. Global Clipping

The existing DP-SGD scheme uses local clipping, i.e., whether or not the *i*-th gradient in the current minibatch is clipped depends only on the *i*-th gradient's norm.

In global clipping, other gradient norms are considered when clipping. It applies the same clipping operation to all gradients within a minibatch.

Clipping Styles

There are two styles of clipping, which can be implemented as either local or global.

- Flat clipping: applies an upper bound to the entire gradient vector, in the form of a norm R
- Layerwise clipping: for each layer r, the gradient vector is bounded by a layer-dependent norm R_r
  - Fixed clipping: a fixed clipping threshold is used for all layers
  - Adaptive clipping: the clipping threshold is dynamic with each layer, allowing for further optimization

Effect of Clipping on NTK and Convergence

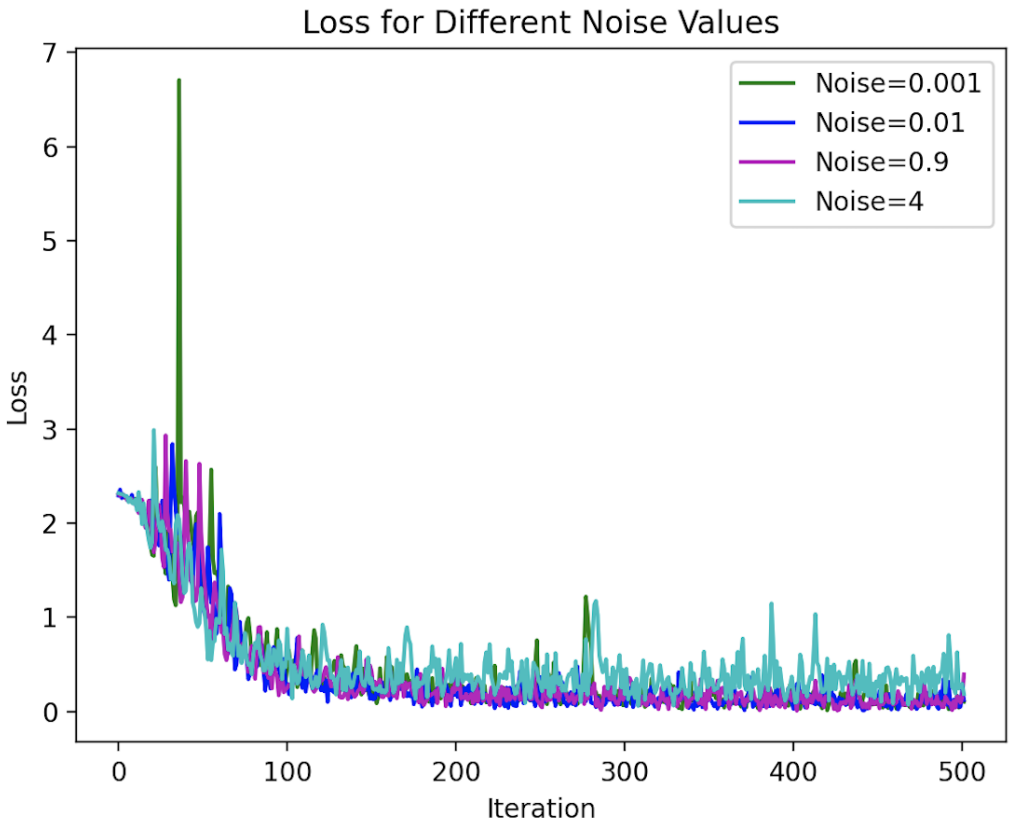
- Local clipping has been shown to break NTK positivity, which worsens convergence.
- Local with flat clipping, preserves positive eigenvalues, results in to 0 convergence.
- Local clipping with both flat and layerwise, loss convergence may not be uniform.
- Global clipping preserves the gradient direction and NTK positivity. This leads to better convergence, while still granting the privacy and computational efficiency achieved in local clipping.

Clipping Types and Convergence

Clipping type	PSD NTK	Loss convergence	To 0 loss
No clipping	Yes	Yes	Yes
Local with Flat	No	No	Yes
Local with Layerwise	No	No	No
Global with Flat	Yes	Yes	Yes
Global with Layerwise	Yes	Yes	Yes

Table: Comparison of different gradient clipping methods

Effect of Noise on Convergence



(a) Convergence of DP-SGD with different noise scales

Figure: CNN trained on MNIST through 500 iterations implementing DP-SGD with local clipping (2 convolutions, 2 linear layers)

Model Overview

To investigate DP-SGD, we trained a logistic regression model on the Fashion-MNIST dataset, using different values for gradient clipping and noise.

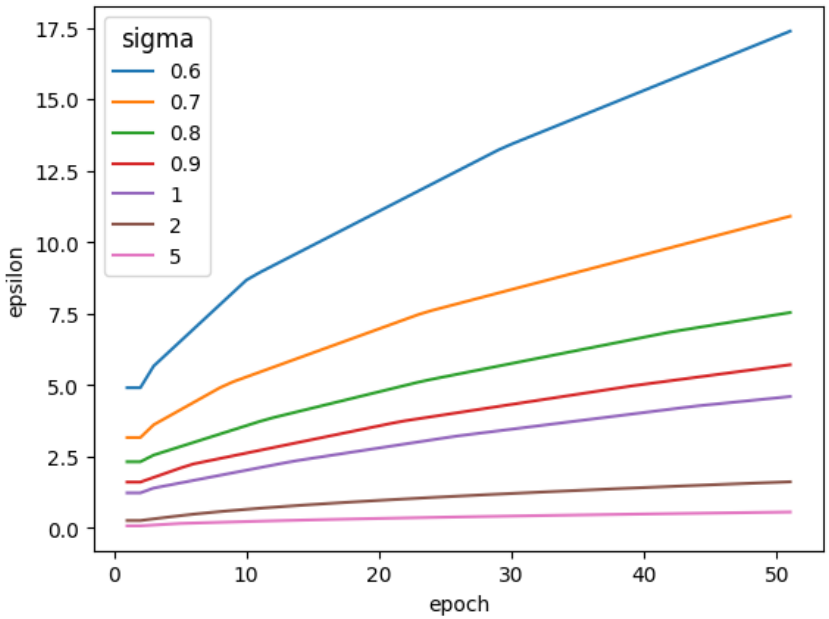
Figure: All Fashion-MNIST Classes

Model Overview

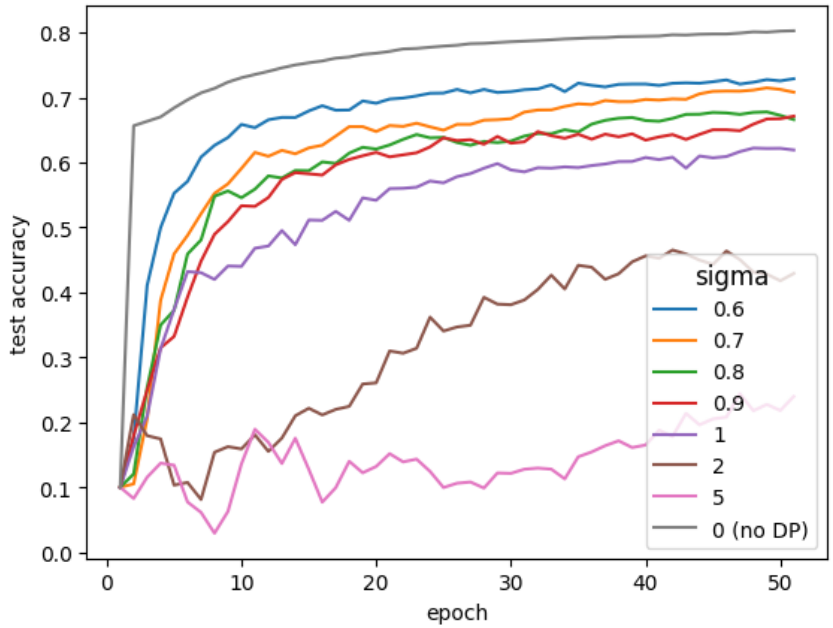
Rényi Differential Privacy (RDP) is a privacy accounting method that is a generalization of ε-differential privacy, and is similar to the moments accountant.

- This technique was used to track the overall privacy loss of training the model.

Experimental Results



(a) Privacy cost for different noise amounts



(b) Test accuracy for different noise amounts

Figure: Privacy cost and test accuracy for different noise amounts on Fashion-MNIST ($\delta = 10^{-5}, C = 0.5$)

Accuracy-Privacy Trade-off

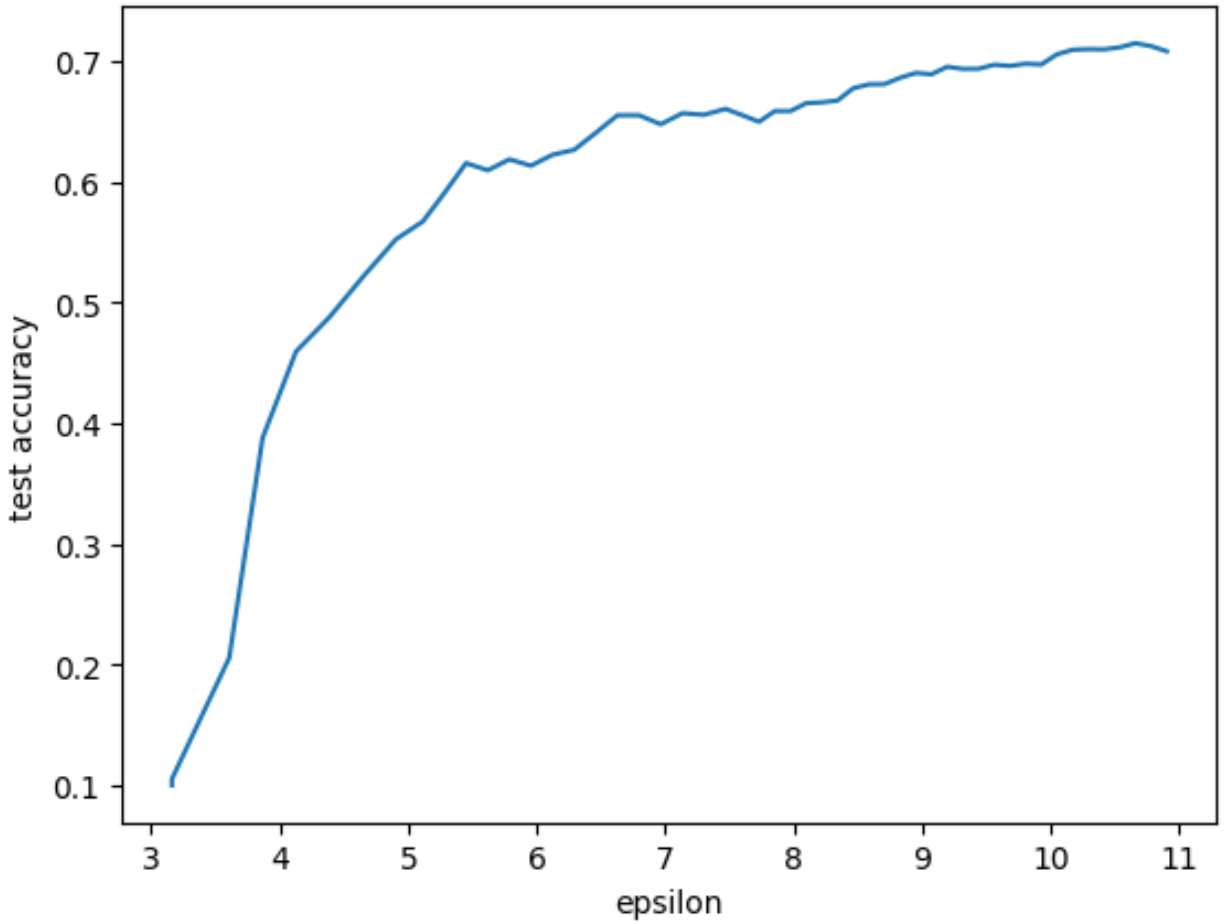
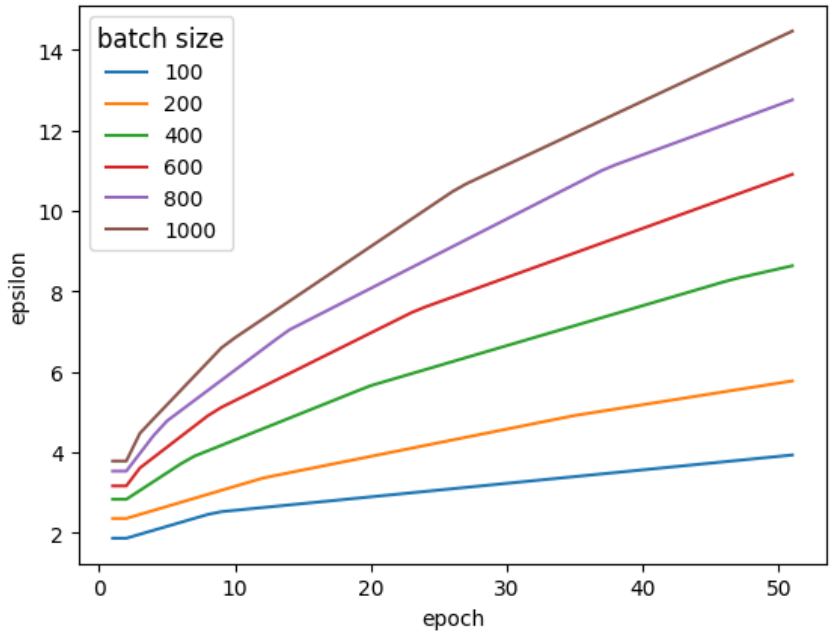
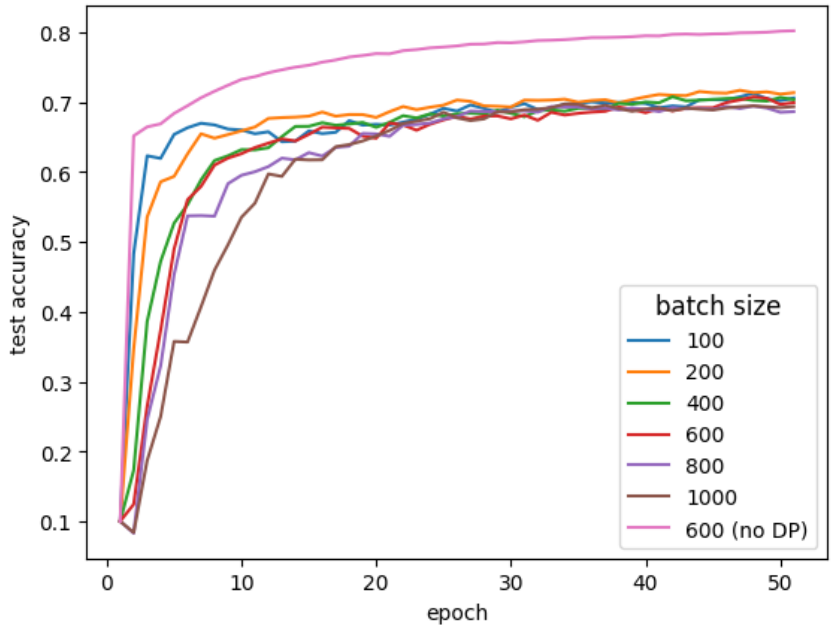


Figure: Test accuracy vs. privacy throughout training on Fashion-MNIST ($\delta = 10^{-5}, \sigma = 0.7, C = 0.5$)

Effects of Batch Size



(a) Batch size vs. privacy



(b) Batch size vs. test accuracy

Figure: Privacy cost and test accuracy for different batch sizes on Fashion-MNIST ($\delta = 10^{-5}$, $\sigma = 0.7$, $C = 0.5$)

Drawbacks of DP-SGD

1. Training with DP-SGD significantly affects utility due to the privacy-accuracy trade-off.
   - Due to noise addition and clipping, DP-SGD is less accurate than non-DP approaches.
2. DP-SGD tends to be very slow in comparison to non-DP optimizers.
   - Per-sample gradient norms are very expensive to compute, leading to high training time.

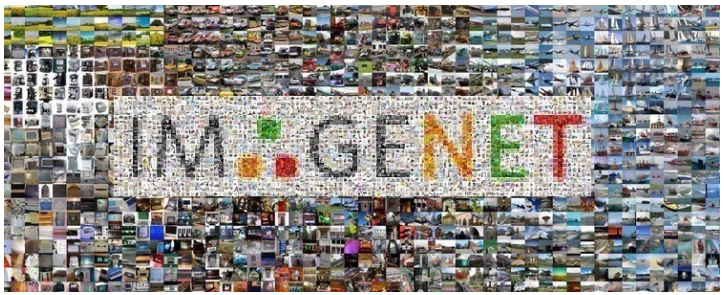
New approaches aim to remedy these drawbacks.

Transfer Learning

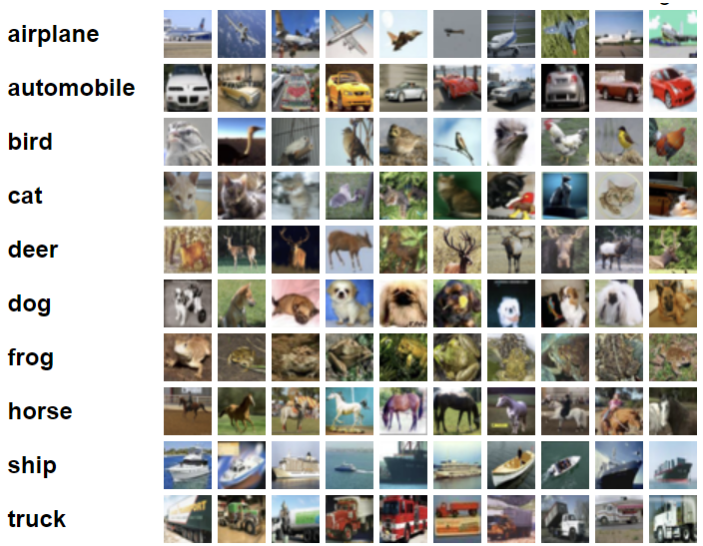
One solution to improve accuracy while minimizing computational cost is transfer learning, in which data is pretrained on public data. This can improve privacy as the model requires less data to reach the same accuracy.

Effect of Transfer Learning on Accuracy

We investigated training a model on the CIFAR-10 datset, using the pretrained ImageNet layer weights. The ImageNet output and input layers were placed with new layers for use on the CIFAR-10 dataset. The hidden layers were frozen.



(a) ImageNet



(b) CIFAR-10

Figure: ImageNet and CIFAR-10 datasets

Transfer Learning Results

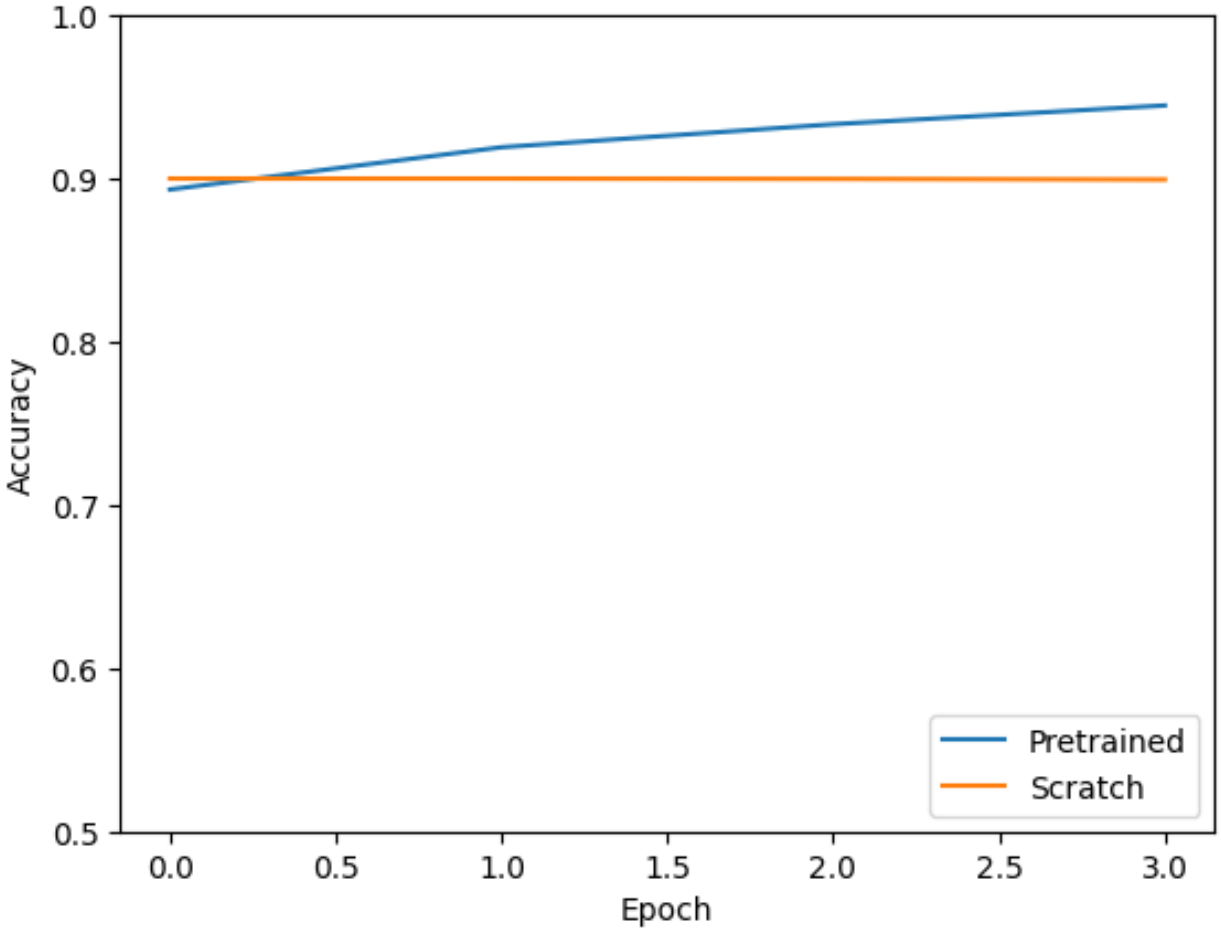


Figure: Pretrained: Transfer learning CIFAR-10 starting with ImageNet weights; scratch model: Dense Keras MaxPooling2D and Conv2D Neural Network (both no DP)

Improving Efficiency

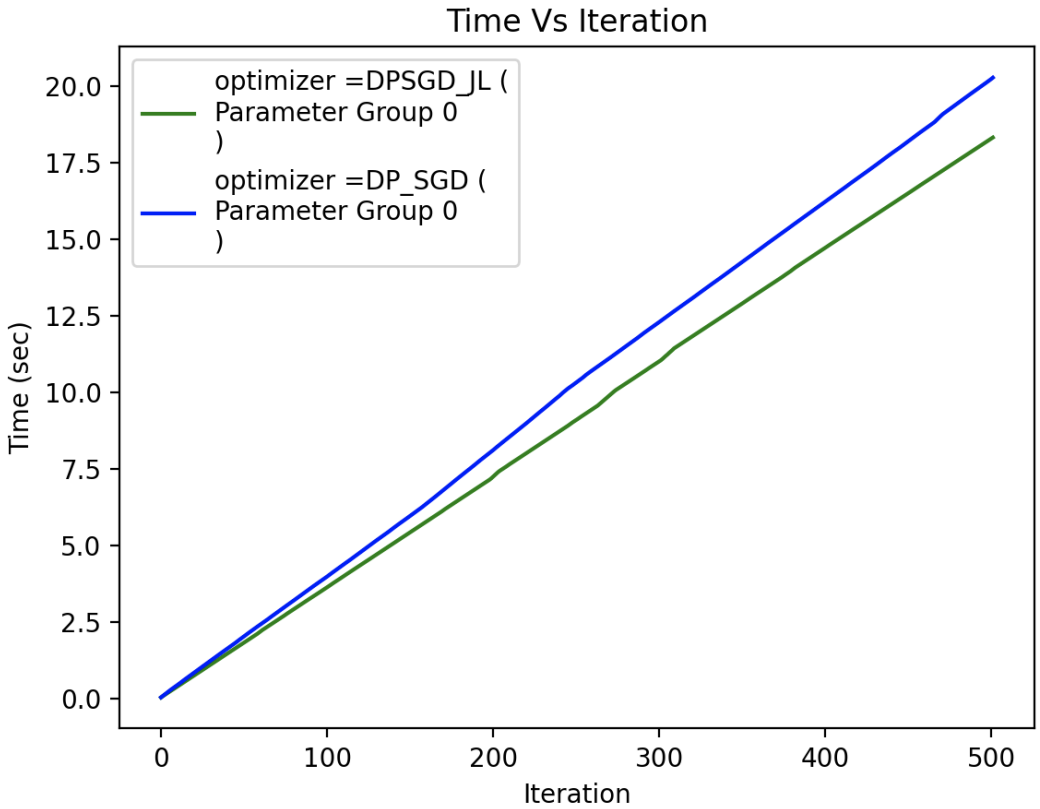
Microbatching: an optimization technique that splits each batch into many "microbatches," which are then clipped

- Gradients are clipped at the microbatch level, rather than individually
- Used in popular libraries such as Tensorflow

DP-SGD-JL: uses Johnson Lindenstrauss (JL) projections to estimate gradient norms rather than fully computing them

- New algorithm developed in 2021
- Results in faster training time and smaller memory footprint

DP-SGD-JL



(a) Speed of DP-SGD vs. DP-SGD-JL

Figure: CNN trained on MNIST through 500 iterations implementing DP-SGD with local clipping (2 convolutions, 2 linear layers, η = 0.1, $\varepsilon = 4$)

References I

- Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. Published online 2016. https://doi.org/10.1145/2976749.2978318
- Applying Differential Privacy to Large Scale Image Classification. ai.googleblog.com. Accessed April 17, 2023. https://ai.googleblog.com/2022/02/applying-differential-privacy-to-large.html
- Bagdasaryan E, Shmatikov V. Differential Privacy Has Disparate Impact on Model Accuracy. arXiv:190512101 [cs, stat]. Published online October 26, 2019. Accessed April 7, 2023. https://arxiv.org/abs/1905.12101
- Bu Z, Gopi S, Kulkarni J, Lee YT, Shen JH, Tantipongpipat U. Fast and Memory Efficient Differentially Private-SGD via JL Projections. arXiv:210203013 [cs]. Published online February 5, 2021. https://arxiv.org/abs/2102.03013
- Bu Z, Wang H, Long Q. On the Convergence and Calibration of Deep Learning with Differential Privacy. arXiv:210607830 [cs, stat]. Published online February 1, 2022. Accessed April 7, 2023. https://arxiv.org/abs/2106.07830

References II

- Differential Privacy - Programming Differential Privacy. programming-dp.com. Accessed April 7, 2023. https://programming-dp.com/ch3.html
- Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Foundations and Trends® in Theoretical Computer Science. 2013;9(3-4):211-407. https://doi.org/10.1561/0400000042
- He J, Li X, Yu D, et al. Exploring the Limits of Differentially Private Deep Learning with Group-wise Clipping. arXiv.org. Published December 3, 2022. Accessed April 17, 2023. https://arxiv.org/abs/2212.01539
- Lundmark M, Dahlman CJ. Differential Privacy and Machine Learning: Calculating Sensitivity with Generated Data Sets Differential Privacy Och Maskininlärning: Beräkning Av Sensitivitet Med Genererade Dataset.; 2017. https://kth.diva-portal.org/smash/get/diva2:1112478/FULLTEXT01.pdf
- Martin S. What Is Transfer Learning? — NVIDIA Blog. The Official NVIDIA Blog. Published February 7, 2019. https://blogs.nvidia.com/blog/2019/02/07/what-is-transfer-learning/

References III

- Rathi M. Deep Learning with Differential Privacy (DP-SGD Explained). mukulrathi.com. https://mukulrathi.com/privacy-preserving-machine-learning/deep-learning-differential-privacy/
- Ridout D, Judd K. Convergence properties of gradient descent noise reduction. Physica D: Nonlinear Phenomena. 2002;165(1-2):26-47. https://doi.org/10.1016/s0167-2789(02)00376-7
- Ruder S. An Overview of Gradient Descent Optimization Algorithms. https://arxiv.org/pdf/1609.04747.pdf
- The Theory of Reconstruction Attacks. differentialprivacy.org. Accessed April 7, 2023. https://differentialprivacy.org/reconstruction-theory/
- Wang YX, Balle B, Kasiviswanathan S. Subsampled Rényi Differential Privacy and Analytical Moments Accountant. Journal of Privacy and Confidentiality. 2020;10(2).

References IV


Fashion MNIST Classes Image, https://www.kaggle.com/code/texasdave/image-classification-tutorial-with-mnist-fashion
https://doi.org/10.29012/jpc.723