Revisiting Benchmarks for Privacy-Preserving Image Classification

by

Sabrina Mokhtari

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

© Sabrina Mokhtari 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is from the following paper, conducted in collaboration with Sara Kodeiri, MMath student at University of Waterloo and Shubhankar Mohapatra, Ph.D student at University of Waterloo, Prof. Florian Tramèr, faculty at the ETH Zürich and my supervisor Prof. Gautam Kamath faculty at University of Waterloo (At the time of writing this thesis, the work is pending peer review.):

- **Sabrina Mokhtari**, Sara Kodeiri, Shubhankar Mohapatra, Florian Tramèr, Gautam Kamath: *Rethinking Benchmarks for Differentially Private Image Classification.* NeurIPS 2024 Datasets and Benchmarks Track [1]

This thesis is a significantly revised and expanded version of the initial paper. In terms of contributions, Mokhtari conducted and executed all experiments related to "No Public Data Pretraining", including ScatterNets and Synthetic Data Pre-training. Mokhtari also developed and implemented the leaderboard and designed and performed the ablation study for "No Public Data Pretraining". Kodeiri managed the experiments inspired by De et al. [24], while Mohapatra focused on the CLIP ViT experiments and CLIP ablation studies. Mokhtari was also responsible for analyzing and writing the experimental results.

[1] Authors SMok, SK, and SMoh are listed alphabetically in order of *first* name. Authors FT and GK are listed in reverse alphabetical order.

Abstract

Differential privacy (DP) is a standard method for preserving the privacy of individual data points. DP prevents models from memorizing training data, thus reducing the risk of data leakage. While DP has been effective in machine learning (ML), there are growing concerns about some common practices in differentially private machine learning (DP ML), particularly the reliance on non-private ML benchmarks to measure progress. Popular datasets like CIFAR-10, while extensively used in non-private settings, may not accurately capture the complexities of privacy-sensitive areas like medical data. Additionally, pre-training on publicly available datasets may not yield the same benefits when the private data differs significantly and is not well represented in the public domain. This thesis addresses these concerns by evaluating DP methods using various privacy-sensitive datasets and training scenarios. We focus on medical datasets, where privacy is crucial, and study a thorough set of techniques. These techniques cover a wide range of settings, including those with public data pre-training, cases without public data, full-layer and last-layer fine-tuning, and different privacy levels.

Acknowledgements

First and foremost, I want to thank Professor Gautam Kamath for sharing his knowledge and always being there to help. Thank you for introducing me to the world of differential privacy and welcoming me into a community I feel incredibly fortunate to be a part of. Your encouragement to participate, ask questions, and have confidence has been invaluable. You've helped me find my voice within the community. What I've learned goes beyond just work; I experienced a very welcoming environment where I felt heard and respected. I am thankful for everything you taught me, and the incredible people I got to meet thanks to you. These are just a small fraction of what I am grateful for and have learned from you. Thank you for this incredible journey.

Furthermore, I want to thank Professor Florian Tramèr for guiding us through our work, sharing his valuable insights, providing great instructive feedback, and helping us along the way with our research questions.

I would like to express my sincere gratitude to Professor Yaoliang Yu and Professor Xi He for serving as the readers of this thesis. I greatly appreciate the time and effort they dedicated to reviewing my work and offering valuable insights. I am also deeply thankful for their presence at my presentation, as their feedback and questions helped strengthen the overall quality of this research.

I want to extend my heartfelt thanks to my family for enduring my absence and offering unwavering support and love throughout. Your belief in me, encouragement to step out of my comfort zone, and inspiration to dream bigger have been invaluable.

Mom, thank you for always being there for me—calming me down when I was upset, encouraging me to move forward when I felt hopeless, and being the most supportive mother I could ever ask for. Thank you for listening to my never-ending stories without ever getting tired. You have been a true friend to me, and I'm so grateful for everything you have done for me and all the sacrifices you have made.

Dad, thank you for being my rock and always being someone I can count on. Your unwavering support gives me the confidence to believe I can achieve anything. Thanks for staying on the call during my deadlines so I would not feel alone and help me even when I don't ask, and for reaching out to everyone you know to ensure I have support wherever I go. They all mean the world to me.

Masih, thank you for being patient with my absence and forgiving me for not being there to watch you grow into the wonderful person you're becoming. Thank you for reminding me every day of how much you love your big sister. Your beautiful and innocent soul has brought so much joy into my life. I love you, my little dada.

I want to extend my gratitude to my friends and family here. First, Khaleh Mahboubeh, thank you for being my family away from home and for your constant support. Your hospitality has been a lifeline. Maryam, thank you for being such a supportive friend and companion through the most difficult times. I also want to thank Fatemeh K. for your warm hospitality and unwavering support—you've been the big sister I always wished for. And to Zahra, thank you for being a joyful and bright soul who always brings laughter! I would like to extend my heartfelt thanks to my teammates, Sara and Shubhankar. I've learned so much from both of you, and it has been an incredible experience working together. I also want to thank everyone I may have missed to mention—thank you all for your support.

Dedication

I dedicate this to my beloved mom, dad, and my sweet little brother. Zohreh, Hossein, and Masih—my three musketeers, I am where I am today because you have always believed in me and supported me through the hardest of times when no one else was around.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vii

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Overview 1

1.2 Thesis Contributions 3

1.3 Thesis Layout 3

2 Background and Problem Statement 4

2.1 Differential Privacy 4

2.1.1 Pure & Approximate Differential Privacy 4

2.1.2 Laplace and Gaussian Mechanism 5

2.1.3 Privacy Loss Random Variable . . . 6
2.1.4 Composition and Post-Processing . . . 6
2.1.5 Privacy Accounting . . . 7
2.1.6 Differentially Private Machine Learning . . . 8
2.2 Problem Statement . . . 8
2.2.1 Common Benchmarks in DP ML . . . 8
2.2.2 Similarity Between Public and Private Data . . . 9
2.2.3 Privacy Considerations in Using Public Data Pre-training . . . 10

3 Related Work **11**
3.1 Differential Privacy in Machine Learning: Approaches, Challenges, and Techniques . . . 11
3.1.1 Training Without Public Data . . . 13
3.1.2 Leveraging Public Data for DP Training . . . 14
3.1.3 Zero-Shot Learning in Differential Privacy . . . 16

4 Dataset and Benchmark Design **17**
4.1 Datasets . . . 17
4.1.1 Private Datasets . . . 18
4.1.2 Public Datasets . . . 19
4.2 Methods and Architectures . . . 20
4.2.1 Overview of Architectures and Techniques . . . 20
4.3 Privacy Parameters . . . 23
4.4 Leaderboard . . . 23

5 Experiments and Results **25**
5.1 Experimental Setup . . . 26
5.2 Revisiting DP ablations . . . 26
5.3 Experimental Findings . . . 29

6 Conclusion and Future Work **34**
6.1 Summary of Thesis and Contributions 34
6.2 Future Work 36
6.2.1 More Datasets and Architectures 36
6.2.2 Language Models 37

References **38**

APPENDICES **49**

A Ablations Study **50**
A.1 Normalization 51
A.1.1 CLIP ViT Architectures 51
A.1.2 ScatterNets 51
A.2 Augmentation Multiplicity 52
A.3 All Ablations Together 52
A.3.1 CIFAR-10 54
A.3.2 CheXpert 54
A.3.3 EyePACS 54

B Additional Details **55**
B.1 Hardware Details 55
B.2 Dataset License 55
B.3 Code & Leaderboard 55

List of Figures

2.1 An overview of differential privacy from [89]. DP ensures the output distribution of a randomized mechanism is similar for two neighboring datasets—datasets that differ by just one data point. This similarity is quantified by a parameter called ε. 5

2.2 Data Visualization. ImageNet [25] and CIFAR-10 [52], commonly used as public-private data pairs in DP pre-training and fine-tuning respectively, are qualitatively similar to each other but different from more privacy-sensitive domains such as EyePACS [36] and CheXpert [39]. 9

4.1 Synthetic data from random processes, StyleGAN [3]. 21

4.2 Overview of CLIP architecture from [73], which is a neural network trained on (image, text) pairs. 22

5.1 Normalization generally improves the final performance of all models. The yellow bars correspond to groups of 8 and 9, while the gray bars correspond to groups of 16 and 27 for CLIP-ViTs and ScatterNet, respectively. The choice of 27 over 81 is due to its superior performance over 9. 27

5.2 Evaluating augmentation multiplicity by adding 4 and 8 augmentations of each image in the training data. Augmentation multiplicity helps in general for CheXpert but not for EyePACS. 28

5.3 Pre-training datasets have varying impacts: Wide-ResNet pre-trained on ImageNet benefits EyePACS most, while ViT-G/14 with linear zero-shot significantly improves CheXpert performance from random guessing, outperforming Wide-ResNet in linear fine-tuning across different ε values and approaching fully fine-tuned Wide-ResNet performance. 30

5.4 Pre-training public data is more beneficial with higher ε values. For CheXpert, ScatterNet performs better at smaller ε values, while pre-trained models show marginal improvements at larger ε values. Similarly, for EyePACS, CLIP ViT-G/14 linear performs better as ε value increases. 31

List of Tables

3.1 Test Accuracy for MNIST and CIFAR-10 under various Differential Privacy (ε-DP) regimes without public data. 13

3.2 Test accuracy for ImageNet and CIFAR-10 datasets under various Differential Privacy (ε-DP) regimes with access to public data. 15

5.1 Test AUC evaluation for EyePACS at different privacy regimes. Baselines include ScatterNet (SN), WideResNet-28-10 (WRN), WideResNet-16-4 (WRN-16-4), and CLIP models on datasets with different public data pretraining. The SOTA is due to [88]. 33

5.2 Test AUC evaluation for CheXpert at different privacy regimes. Baselines include ScatterNet (SN), WideResNet-28-10 (WRN), WideResNet-16-4 (WRN-16-4), and CLIP models on datasets with different public data pretraining. The SOTA is due to [7](Private) and [98](Non-Private). 33

A.1 Studying the impact of normalization for CLIP ViT models, we find that normalization generally improves performance, but the effectiveness depends on the architecture and dataset. Normalizations marked in red indicate a drop in performance compared to the baseline. 51

A.2 Studying the impact of normalization for ScatterNet + CNN, we observe that normalization consistently improves performance. Data normalization tends to outperform group normalization for EyePACS and CIFAR-10, particularly due to the large gradients of their Scatter features. 51

A.3 Studying the impact of augmentation multiplicity, we find that it consistently improves performance for CIFAR-10. However, looking at EyePACS and CheXpert, we observe inconsistent behavior, except that it generally seems to reduce performance with EyePACS. "Best so far" refers to the baseline plus all previously effective ablations listed in Table A.4 52

A.4 Studying all ablations together, we observe an almost consistent improvement in performance for CIFAR-10, whereas this pattern is not observed with the other datasets. 53
A.5 Studying all ablations for CIFAR-10. 54
A.6 Studying all ablations for CheXpert. 54
A.7 Studying all ablations for EyePACS. 54

Chapter 1

Introduction

1.1 Overview

Machine learning (ML) models memorize training data, making them susceptible to various types of data leakage attacks, including membership inference attacks [42, 77], model inversion[37], and training data extraction [16, 17, 78]. These vulnerabilities are especially concerning when the training data contains sensitive information. Ensuring the privacy of training data is therefore a critical challenge in the deployment of ML models.

Differential privacy (DP) [30] is the standard approach for protecting the privacy of individual data points in statistical analysis and machine learning. It ensures privacy by preventing ML models from memorizing their training data, thus avoiding the leakage of sensitive information. Differentially Private Stochastic Gradient Descent (DP-SGD) is the key technique for training differentially private machine learning (DP ML) models, adding controlled noise to gradients during training to limit the influence of each data point. It has widely been used for various applications such as training private image classifiers [1, 86, 67, 24], large language models [2, 94, 57], and other generative models [91, 6, 15, 8, 26, 40]. However, this protection often comes at the cost of model utility. To address this shortcoming, extensive research has focused on enhancing DP learning algorithms with pre-training on public data [95, 68, 33, 50, 24, 7].

Recently, Tramèr, Kamath, and Carlini [87] have raised critical concerns regarding some common trends in DP ML. Much of the existing research has relied on canonical datasets such as MNIST [55], CIFAR-10 [52], and ImageNet [25], which are widely used in non-private ML settings. Although these datasets have some privacy concerns [9], their

relevance to domains where privacy is of utmost importance, such as medical data, remains questionable. These datasets primarily consist of natural images or handwritten digits, which do not adequately represent the complexities and sensitivities inherent in more privacy-sensitive domains.

Additionally, they point out that many existing techniques in the literature often involve pre-training with public data that is qualitatively similar to the private data used for fine-tuning. A common scenario is pre-training on ImageNet and then fine-tuning on CIFAR-10, both containing very similar natural images. While this method has its advantages, it may not accurately represent the challenges encountered in real-world applications where private data is not well-represented by public datasets. Consequently, it is unclear how much these common public datasets are effective in settings where the private data is qualitatively different.

Additionally, they raise concerns about the sensitive information present in public datasets. Many public datasets may contain sensitive content related to individuals. For example, despite its widespread use, ImageNet contains sensitive personal content [72]. TinyImages [85] is another example of a dataset that was later deprecated due to containing offensive content [9]. Originally consisting of 80 million images scraped from the web, TinyImages was subsampled to create CIFAR-10. It is noteworthy that TinyImages has since been deprecated [1]. The challenges are even greater for large-scale datasets used in natural language processing (NLP). These datasets are typically curated by scraping the internet with minimal filtering or oversight.

These critiques provide a fresh perspective on some of the common trends in DP ML and raise significant questions:

- "Are we truly measuring progress in DP ML?"
- "Does success on widely adopted datasets translate to those that are privacy-sensitive?"
- "How applicable are current DP ML methods and trends when we shift towards more meaningful datasets?"

In this thesis, these questions are addressed, particularly in the field of image classification—a prominent area of research—by focusing on more privacy-sensitive domains, i.e., medical datasets.

[1] https://groups.csail.mit.edu/vision/TinyImages/

1.2 Thesis Contributions

We conduct extensive experiments with various architectures, training regimes, privacy levels, and pre-training and fine-tuning data pairs. Our goal is to evaluate the validity of the concerns raised by Tramèr, Kamath, and Carlini [87] and to provide insights into how progress in DP ML can be better measured and applied to privacy-sensitive datasets. Our contributions are as follows:

- We propose standardized benchmark datasets and evaluation settings to measure progress in DP image classification, with a particular focus on privacy-sensitive domains.
- We release a public leaderboard for DP ML, for the community to track improvements on these benchmarks.
- We evaluate previously established techniques for DP image classification across a variety of settings to see which are and are not broadly effective.

1.3 Thesis Layout

This thesis is organized as follows. Chapter 2 introduces preliminaries and foundational concepts related to DP in general, and in ML. This chapter also provides a detailed explanation of the problem definition. Chapter 3 presents a literature review, focusing on common trends in DP ML, particularly in image classification. We discuss the datasets used for both fine-tuning and pre-training in previous works, the methods applied, and the results achieved. Chapter 4 focuses on datasets and benchmarks design, outlining the methods and datasets used in our experiments. We explain our approach and the rationale behind the choices made for the benchmarks. Chapter 5 presents and analyzes the results of our experiments, detailing our findings. Finally, Chapter 6 summarizes the thesis, highlights our contributions, and suggests potential areas for future work.

Chapter 2

Background and Problem Statement

In this Chapter, we provide an overview of differential privacy (DP), preliminary knowledge, foundational concepts, and various methods used in both theoretical DP and practical applications. We also discuss DP in the context of machine learning (ML) and offer an in-depth examination of the problem statement.

2.1 Differential Privacy

For two neighboring datasets—datasets that differ by only one data point—ε defines the maximum distance between a random mechanism's output distribution when applied to these two datasets as shown in Figure 2.1. In the context of tabular data, neighboring datasets differ by a single row, whereas for batches of images, they differ by one image. This ensures that the algorithm's output remains rather unchanged regardless of the inclusion or exclusion of any single data point, thereby preserving individual privacy.

2.1.1 Pure & Approximate Differential Privacy

A randomized algorithm M achieves ε-DP if for all $Z \subseteq \text{Range}(\mathcal{M})$ and for any two neighboring datasets $D, D' \in \mathcal{D}$ that differ in one data point:

$$\Pr[\mathcal{M}(D) \in Z] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in Z].$$

The above notion is the definition of pure DP. Approx DP is a commonly used relaxation of pure DP (ε-DP) and is in the form of (ε, δ)-DP [31, 29]. A randomized algorithm M

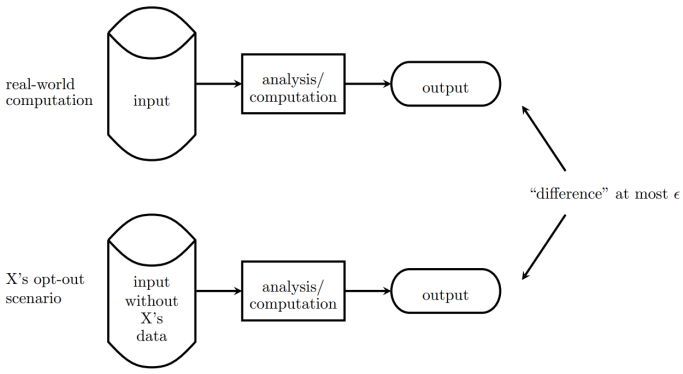


Figure 2.1: An overview of differential privacy from [89]. DP ensures the output distribution of a randomized mechanism is similar for two neighboring datasets—datasets that differ by just one data point. This similarity is quantified by a parameter called ε.

achieves (ε, δ)-DP if for all $Z \subseteq \text{Range}(\mathcal{M})$ and for any two neighboring datasets $D, D' \in \mathcal{D}$ that differ in one data point:

$$\Pr[\mathcal{M}(D) \in Z] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in Z] + \delta.$$

The privacy cost is measured by the parameters (ε, δ), often referred to also as the privacy budget. Smaller privacy parameters suggest stronger privacy. When $\delta = 0$, we call the privacy measure pure DP.

2.1.2 Laplace and Gaussian Mechanism

Gaussian mechanism [31] and Laplace mechanism [30] are two widely used DP algorithms that add noise to the random mechanism. Given a function $f : \mathcal{D} \to \mathbb{R}^d$, the Gaussian mechanism adds noise sampled from a Gaussian distribution $\mathcal{N}(0, S_f^2\sigma^2)$ to each component of the query output, where σ is the noise scale and S_f is the L_2 sensitivity of function f, which is defined as $S_f = \max_{D,D' \text{differ in one data point}} ||f(D) - f(D')||_2$. For $\varepsilon \in (0, 1)$, if $\sigma \geq \sqrt{2\ln(1.25/\delta)}/\varepsilon$, the Gaussian mechanism satisfies (ε, δ)-DP.

Laplace mechanism works similarly but with the noise from the Laplace distribution and the L_1 sensitivity.

2.1.3 Privacy Loss Random Variable

Let Y and Z be two random variables. The privacy loss random variable $\mathcal{L}_{Y\|Z}$ is distributed by drawing $t \sim Y$, and outputting $\ln\left(\frac{\Pr[Y=t]}{\Pr[Z=t]}\right)$. If the supports of Y and Z are not equal, then the privacy loss random variable is undefined [46].

This concept applies to general random variables but is usually used with Y and Z set to $M(X)$ and $M(X')$, where M is an algorithm and X and X' are neighboring datasets. Essentially, the privacy loss random variable helps determine the likelihood that X was used as the input dataset instead of X', based on the observed output of the algorithm, $M(X)$. This measures how the presence or absence of a single data point in the dataset affects the algorithm's output, indicating the level of privacy maintained.

The definition of pure differential privacy indicates that ε-DP means the absolute value of the privacy loss random variable, denoted as $\left|\mathcal{L}_{M(X)|M(X')}\right|$, is bounded by ε for any neighboring datasets X and X'. In other words, ε-DP ensures that with probability 1, the privacy loss is no greater than ε.

For (ε, δ)-DP, the definition is slightly relaxed. It implies that the absolute value of the privacy loss random variable is bounded by ε with probability at least $1 - \delta$. This allows for a small probability δ where the privacy loss might exceed ε, providing a balance between privacy and utility.

2.1.4 Composition and Post-Processing

Complex DP algorithms can be built from these basic algorithms following two important properties of DP:

1. Post-processing [29] states that for any function g defined over the output of the mechanism $\mathcal{M}$, if $\mathcal{M}$ satisfies (ε, δ)-DP, so does $g(\mathcal{M})$.

2. Composition [28] states that if $M = (M_1, ..., M_k)$ and $\mathcal{M}_1$, $\mathcal{M}_2$, $\cdots$, $\mathcal{M}_k$ satisfy $(\varepsilon_1, \delta_1)$-, $(\varepsilon_2, \delta_2)$-, $\cdots$, $(\varepsilon_k, \delta_k)$-DP, then if M_i's are potentially chosen sequentially and adaptively, M is $(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i)$-differentially private.

3. Advanced Composition [32] states that if all $\varepsilon_i = \varepsilon$, and all $\delta_i = \delta$, then M will overall be $\left(\varepsilon\sqrt{8k\ln(1/\delta')}, k\delta + \delta'\right)$-DP. This method results in an $O(\sqrt{k})$ increase in the value of ε, unlike basic composition, which results in an increase by a factor of k.

After training ML models, the post-processing property ensures that the model can be released publicly without additional privacy concerns. Meanwhile, composition allows for iterative private training processes, determining the final privacy budget spent.

2.1.5 Privacy Accounting

All DP algorithms have a privacy proof that provides an upper bound on the values of ε and δ if approximate DP is used. For DP-SGD, this is typically handled by privacy accountants, which take into account various parameters of the model and training process, such as clipping norm, noise multiplier, and data size, to calculate the final values of these two privacy parameters. There have been various accounting methods used in the literature, but the following two are among the most famous ones.

Rènyi Differential Privacy

Rényi differential privacy (RDP) [59] is a variant of DP. It provides a family of privacy guarantees based on the Rényi divergence, which measures the distance between two probability distributions. In RDP, the privacy guarantee is parameterized by a non-negative real number ε, which controls the amount of privacy protection provided. Additionally, RDP can be useful when incorporating composition properties as it offers tighter privacy bounds. A mechanism M is (α, ε)-RDP with $\alpha \in (1, \infty)$ if for all neighboring datasets $D, D^{'}$, $D_{\alpha}\left(\mathcal{M}(X)|\mathcal{M}\left(D'\right)\right) = \frac{1}{\alpha-1}\log E_{\theta\sim D}\left[\left(\frac{p\mathcal{M}(D)(\theta)}{p\mathcal{M}(D')(\theta)}\right)^{\alpha}\right] \leq \varepsilon$

If a mechanism M_1 satisfies (α, ε)-RDP, then M_1 also satisfies $(\varepsilon + \frac{log1/\delta}{\alpha-1}, \delta)$-DP for any $\delta \in (0, 1)$.

Let $M = (M_1, ..., M_n)$ be mechanisms where M_k can potentially depend on the outputs of $M_1, ..., M_{k-1}$. Then M obeys RDP with $\varepsilon_{\mathcal{M}_1\times\mathcal{M}_2}(\cdot) = \left[\varepsilon_{\mathcal{M}_1} + \varepsilon_{\mathcal{M}_2}\right](\cdot)$.

Privacy Random Variable Accountant

Privacy Random Variable (PRV) Accountant is a fast and efficient algorithm for composing privacy guarantees of DP algorithms with arbitrary accuracy [38]. The key innovation lies in the use of privacy loss random variables to quantify privacy loss. The algorithm achieves a running time and memory complexity of $\tilde{O}(\sqrt{k})$ to approximate the privacy curve of a DP algorithm composed with itself k times, which is a substantial improvement over the previous best method that required $\tilde{\Omega}(k^{1.5})$ running time. It provides tighter upper bounds compared to RDP.

2.1.6 Differentially Private Machine Learning

Differentially Private Stochastic Descent

Differentially Private Stochastic Gradient Descent (DP-SGD) [1] is an algorithm used in machine learning that provides differential privacy to the training data. DP-SGD adapts standard SGD to satisfy the constraints of differential privacy. The main difference between this approach and traditional training is that in DP-SGD, gradients are clipped according to $\overline{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$ using a clipping norm C, and then privatized according to $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \overline{\mathbf{g}}_t(x_i) + \mathcal{N}\left(0, \sigma^2 C^2 \mathbf{I}\right)\right)$. If SGD optimizer is replaced by Adam, then the algorithm is called DP-Adam.

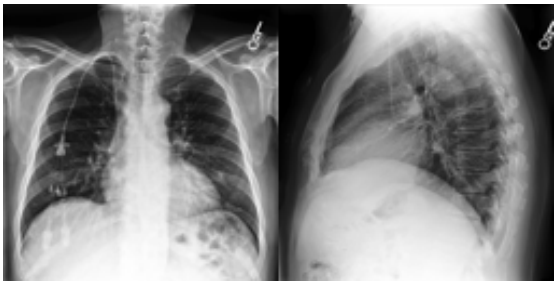
2.2 Problem Statement

Recently, Tramèr, Kamath, and Carlini [87] have raised significant critiques about several trends in differentially private machine learning (DP ML).

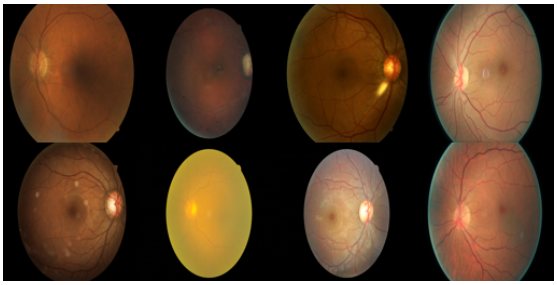
2.2.1 Common Benchmarks in DP ML

One key critique questions whether the benchmarks used in DP ML truly measure progress in the field, particularly in image classification, which is the focus of this thesis. Informally speaking, recent datasets that have been widely adopted in the domain are not any different from those that are used in non-private domains. While significant progress has been made, they question whether this progress generalizes to privacy-sensitive settings where DP may be deployed.

For example, CIFAR-10 and ImageNet are both composed primarily of natural images of everyday objects. While these datasets indeed have some privacy concerns [9], it is less clear whether they resemble domains where privacy is a critical concern, such as medical images. Additionally, some of these datasets are nearly solved, meaning that we can achieve very high accuracy even without additional training, provided there are no restrictions on the pre-training data or methods used. For instance, using the CLIP ViT-G [76] model, we can achieve 99.75% zero-shot accuracy on CIFAR-10. This raises doubts about the relevance of these datasets for evaluating DP ML, especially in settings where there are no limits on public data. Nevertheless, the widespread use of these datasets still allows for direct comparisons of accuracy, which helps track how well differential privacy models perform compared to non-private models.



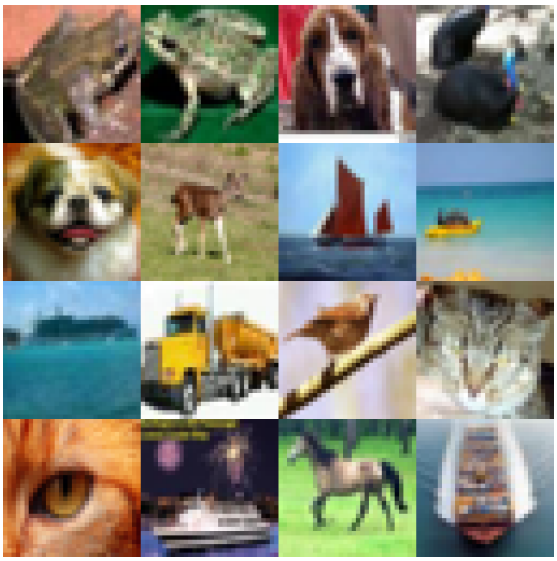
(a) CheXpert Samples



(b) EyePACS Samples



(c) ImageNet Samples



(d) CIFAR-10 Samples

Figure 2.2: Data Visualization. ImageNet [25] and CIFAR-10 [52], commonly used as public-private data pairs in DP pre-training and fine-tuning respectively, are qualitatively similar to each other but different from more privacy-sensitive domains such as EyePACS [36] and CheXpert [39].

2.2.2 Similarity Between Public and Private Data

Tramèr, Kamath, and Carlini [87] also point out that in many cases, the pre-training and private data used in DP ML research seem to come from very similar distributions. A common scenario is pre-training a model on ImageNet and fine-tuning it on CIFAR-10. Informally speaking, they generally contain very similar images; also, CIFAR-10 labels

are already represented in ImageNet as shown in Figure 2.2. This raises the question of whether the model is performing any private learning or if it has already been trained to classify the private data before any fine-tuning takes place. This concern extends to the choice of public data and whether available public datasets can effectively help when dealing with private datasets that are informally speaking significantly different in their distribution or are less represented within public datasets.

2.2.3 Privacy Considerations in Using Public Data Pre-training

Moreover, not all public data should be considered truly public. The web is a vast repository of data, some of which is privacy-sensitive. It is imperative to consider the privacy implications of using web-sourced data, as it may inadvertently include personal or sensitive information. For instance, datasets that are assumed to be public could contain sensitive or inappropriate content [72, 9]. For instance, some of the previously widely used datasets have been deprecated due to containing inappropriate content [84]. Additionally, some datasets that are considered for pre-training and have been used in the literature are proprietary and not widely accessible [7]. This underscores the importance of selecting appropriate public data and also highlights the need to develop methods that do not rely heavily on such datasets. Leveraging synthetic datasets to learn valuable image priors or using handcrafted features to create artificial priors can be beneficial in these contexts. Therefore, a comprehensive approach to DP ML must account for the varying nature of data sources and employ various techniques to take into account all the previous arguments and ensure privacy and effectiveness.

Chapter 3

Related Work

In Chapter 2, we introduced the foundational concepts of differential privacy (DP) and its application in machine learning (ML). We also offered a detailed explanation of the problem statement. In this Chapter, we present prevailing trends and best practices in the domain of differentially private machine learning. We specifically examine scenarios where public data is not used for pre-training and focus on privately training from scratch, exploring what previous research has accomplished in this area. We also investigate recent work that incorporates public data for pre-training, comparing some of the state-of-the-art methods and their performances. Additionally, following the suggestions of Tramèr, Kamath, and Carlini [87], we analyze which datasets have been frequently used in the literature and provide a comprehensive summary of past research.

3.1 Differential Privacy in Machine Learning: Approaches, Challenges, and Techniques

Machine learning (ML) models memorize information from their training data, which makes them susceptible to various types of attacks [42, 77, 37, 16, 17, 78]. Differential privacy (DP) [30, 29] has emerged as a robust approach to mitigate such risks by preserving the privacy of individual data points. It works by limiting the influence any single data point in the training set can have on the model's output, thereby safeguarding their privacy. The level of privacy provided by DP is typically measured with a parameter ε. A smaller ε value indicates stronger privacy protection. This ability of DP has made it be widely adopted in various applications of image classification and large language models [1, 86,

67, 24, 2, 94, 57]. However, differentially private learning scales poorly and significantly sacrifices the model's useful generalization capabilities to provably prevent memorization. Therefore, researchers have focused on enhancing differentially private machine learning (DP ML) performance by employing various techniques such as leveraging public data for pre-training [95, 68, 33, 50, 24, 7], using synthetic data for pre-training [82], or using better handcrafted features [86].

Most previous research in this domain, specifically in the context of differentially private image classification, has focused on standard benchmarks such as MNIST [55], CIFAR-10 [52], and ImageNet [25]. Recently, some new studies extend their work by exploring more privacy-sensitive datasets, particularly medical datasets like CheXpert [39, 7, 74], Google-proprietary DR dataset [74], and MIMIC-CXR [44, 7]. The architectures employed in these studies have primarily involved variations of ResNets [100] and their various derivatives, such as Normalizer-Free ResNets (NF-ResNet) [10]. In addition to ResNet variations, Vision Transformers (ViTs) [27] have been utilized for their ability to handle large-scale data and capture long-range dependencies effectively. They have shown promise in several image classifications capabilities [83].

Linear probes have also been a common approach, focusing on extracting valuable features and incorporating efficient DP fine-tuning methods. Typically, this method involves pre-training a model on a large dataset without privacy and then DP fine-tuning only the last classifier layer on the private dataset. This technique allows researchers to leverage the rich representations learned from public datasets or synthetic data [83]. Additionally, there has been exploration into using handcrafted features combined with linear classifiers. Tramèr and Boneh [86], for example, have investigated this approach by using the Scattering Network (ScatterNet) [63], a non-learned SIFT-like feature extractor that can be used with simple linear classifiers. This method leverages handcrafted features that do not incur any privacy cost, as they are not trained on any dataset.

Other convolutional neural network (CNN) architectures have also been explored in DP ML [86, 68]. The diversification in dataset and architecture choice aims to improve the generalizability and applicability of DP ML across different domains and tasks. By integrating a wide range of datasets and leveraging advanced architectural techniques, researchers strive to push the boundaries of what is achievable with DP ML, ultimately contributing to the development of more robust and privacy-preserving models.

In addition to using various datasets, architectures, and techniques, researchers also study different levels of privacy in their work. They experiment with a range of privacy parameters, from tight privacy bounds (smaller ε values) to looser privacy guarantees (larger ε values). This makes it difficult to directly compare results across different studies.

Next, we will look into several previous studies to identify common trends and datasets explored in the literature.

3.1.1 Training Without Public Data

Private Data	ε-DP	Source	Test Accuracy (%)
MNIST	6.78	Yu et al. [97]	93.2
MNIST	6	Nasr et al. [62]	96.1
MNIST	2.93	Papernot et al. [65]	98.1
MNIST	1.97	Papernot et al. [66]	98.5
MNIST	3	Tramèr and Boneh [86]	**98.7**
CIFAR-10	8	Yu et al [96]	55.9
CIFAR-10	7.53	Papernot et al. [68]	66.2
CIFAR-10	1.93	Dörmann et al.[33]	58.6
CIFAR-10	4.21	Dörmann et al.[33]	66.2
CIFAR-10	7.42	Dörmann et al.[33]	70.1
CIFAR-10	7.5	Klause et al. [50]	71.7
CIFAR-10	3	Tramèr and Boneh [86]	66.9
CIFAR-10	1	De et al. [23]	56.8
CIFAR-10	2	De et al. [23]	65.9
CIFAR-10	4	De et al. [23]	73.5
CIFAR-10	8	De et al. [23]	**81.4**
CIFAR-10	3	Tang et al. [82]	81

Table 3.1: Test Accuracy for MNIST and CIFAR-10 under various Differential Privacy (ε-DP) regimes without public data.

Yu et al. [96] utilize a Wide-ResNet 28-4 architecture, achieving 55.9% accuracy on CIFAR-10 with ε set to 8. Papernot et al.[68] report a 66.2% accuracy with ε 7.53 and δ of 10^{-5} using a custom CNN architecture. Dörmann et al.[33] achieve 58.6%, 66.2%, and 70.1% accuracy on CIFAR-10 with ε values of 1.93, 4.21, and 7.42, respectively, using the same architecture described in [68]. Klause et al. [50] achieve 71.7% with ε 7.5 on CIFAR-10. Additionally, Tramèr and Boneh [86] obtain 65.4%, 66.9%, and 52.5% accuracy for ε 3 using handcrafted features with a linear probe, a custom CNN layer, and an end-to-end CNN, respectively. All these results were achieved with δ set to 10^{-5}.

Following the successful use of synthetic data learned from random processes [56, 14, 35, 47] for representation learning [41, 20, 18] and to get valuable visual priors [48, 3], Tang et al. [82] utilize noise priors from synthetic images to enhance DP training performance. This strategy is advantageous because training on synthetic data incurs no privacy cost due to significant distribution shifts from natural images. They achieve 72.32%, 77.25%, 79.99%, 81.88%, 84.01%, 85.26% for ε values of 1, 2, 3, 4, and 6, respectively, using a Wide-ResNet 16-4 on CIFAR-10, outperforming previous "no public pre-training" regime results on CIFAR-10. Additionally, using only linear probing, they achieve 40.64%, 57.10%, 60.89%, 65.10%, 67.78%, 69.92%, and 71.08% for ε values of 0.03, 0.1, 0.2, 0.5, 1, 2, and 3, respectively, surpassing the results of Tramèr and Boneh's handcrafted features plus linear probe and CNN approach. Moreover, they achieve a 39.39% accuracy at $\varepsilon = 8$, compared to the previous SOTA [75] on ImageNet. They use a Vision Transformers (ViT) [27] model pre-trained on Shaders-21k [96]. A summary of these prior works is presented in Table 3.1.

3.1.2 Leveraging Public Data for DP Training

As mentioned earlier, using DP for privately training ML models can negatively impact performance. Theoretical analysis of Differentially Private Stochastic Gradient Descent (DP-SGD) [1], one of the most prominent techniques for training DP ML models, reveals that adding noise significantly hinders convergence, especially at the beginning of training. This noise disrupts the learning process, making it difficult for the model to identify meaningful patterns in the early stages.

To mitigate this issue, pre-training on public data can greatly enhance initial convergence by providing a better starting point for fine-tuning. Public data, such as ImageNet, offers a wealth of visual priors that can be leveraged to give the model a head start before fine-tuning on private datasets. Previous studies have demonstrated the effectiveness of this strategy [24, 7, 50, 86, 25, 19, 53, 101, 11, 80, 58]. By pre-training models on public datasets, researchers can transfer learned features and representations to private data with minimal additional training. For example, models pre-trained on ImageNet without privacy constraints and subsequently fine-tuned with DP on private datasets like CIFAR-10 have shown substantial improvements in accuracy compared to models trained from scratch with DP constraints. Thus, leveraging public data for pre-training has become a focal point in the research community. Numerous studies have explored different architectures and pre-training techniques to optimize the trade-off between privacy and utility, which we will examine in detail.

Without additional data, De et al. [24] achieve 56.8%, 65.9%, 73.5%, and 81.4% top-1 accuracy on CIFAR-10 for ε values of 1, 2, 4, and 8 respectively, with δ set to 10^{-5}, using

Public Data	Private Data	ε-DP	Source	Test Accuracy (%)
ImageNet	CIFAR-10	1	De et al. [24]	94.7
ImageNet	CIFAR-10	2	De et al. [24]	95.4
ImageNet	CIFAR-10	4	De et al. [24]	96.1
ImageNet	CIFAR-10	8	De et al. [24]	**96.7**
CIFAR-100	CIFAR-10	2	Tramèr and Boneh [86]	80
ImageNet	CIFAR-10	2	Tramèr and Boneh [86]	92.7
Places36	ImageNet	10	Kurakin [53]	47.9
JFT-4B	ImageNet	1	De et al. [24]	84.4
JFT-4B	ImageNet	2	De et al. [24]	85.6
JFT-4B	ImageNet	4	De et al. [24]	86.0
JFT-4B	ImageNet	8	De et al. [24]	**86.7**
JFT-4B	ImageNet	0.25	Mehta et al. [58]	75.6
JFT-4B	ImageNet	0.5	Mehta et al. [58]	79.4
JFT-4B	ImageNet	1	Mehta et al. [58]	81.1
JFT-4B	ImageNet	2	Mehta et al. [58]	81.5
JFT-4B	ImageNet	4, 6, 8, 10	Mehta et al. [58]	81.7

Table 3.2: Test accuracy for ImageNet and CIFAR-10 datasets under various Differential Privacy (ε-DP) regimes with access to public data.

WRN-16-4 and WRN-40-4 models. However, when utilizing pre-trained WRN-28-10 and WRN-40-4 models on ImageNet, they achieve state-of-the-art (SOTA) accuracies of 94.7%, 95.4%, 96.1%, and 96.7% under the same privacy regime using full fine-tuning and also achieve 93.1%, 93.6%, 94%, 94.2% with the same ε values by only fine-tuning the classifier layer. Tramèr and Boneh [86] use ResNeXt [92] pre-trained on labeled CIFAR-100 and achieve 80.0% at a budget of $(\varepsilon = 2, \delta = 10^{-5})$ on CIFAR-10. They further train a ResNet-50 model on unlabeled ImageNet [25] using SimCLRv2 [19] and achieve 92.7% accuracy for a DP budget of $(\varepsilon = 2, \delta = 10^{-5})$.

Kurakin et al. [53] achieve 47.9% accuracy on ImageNet by pre-training a ResNet-18 on Places365 [101] using a ResNet-18 model. Additionally, without using public data, De et al. [24] achieve 32.4% top-1 accuracy on ImageNet with $\varepsilon = 8$ and $\delta = 8 \times 10^{-5}$ using the NF-ResNet-50 model [11], a modified version of ResNet-50 designed to enhance signal propagation. However, they achieve SOTA results when the model is pre-trained on JFT-4B, Google's proprietary dataset containing 4 billion images from 30k classes [80]. They achieve top-1 accuracies of 84.4%, 85.6%, 86.0%, and 86.7% with ε values of 1, 2, 4, and

8, respectively, and $\delta = 8 \times 10^{-5}$. Mehta et al. [58] pre-train ViT-H/14-4b model on JFT and finetune on ImageNet and achieve 75.6%, 79.4%, 81.1%, 81.5%, 81.7%, 81.7%, 81.7%, 81.7% for ε of 0.25, 0.5, 1, 2, 4, 6, 8, 10 and $\delta = 10^{-6}$ respectively. A summary of these prior works is presented in Table 3.2.

3.1.3 Zero-Shot Learning in Differential Privacy

As previously noted, zero-shot learning is an important area of research in both privacy and non-privacy settings. This approach is particularly interesting in private ML because it effectively translates to an ε of 0 (perfect privacy) in the DP framework. This means there is no privacy risk, as the model is never directly trained on the private dataset. One notable example is OpenAI's CLIP models [73], which have demonstrated strong zero-shot capabilities when pre-trained on large datasets from the Internet. Radford et al. [73] achieve a 76.2% zero-shot accuracy on ImageNet and over 90% zero-shot accuracy on CIFAR-10.

Chapter 4

Dataset and Benchmark Design

In Chapter 3, we conducted a literature review, examining which datasets have been considered private and used for evaluation in previous research. We explored the most successful approaches in differentially private machine learning (DP ML), with a particular focus on public data pre-training. We observed that much of the research has centered around canonical datasets such as MNIST, CIFAR-10, and ImageNet. We noted that public-private dataset pairs often share similar distributions, with many studies using ImageNet for pre-training and CIFAR-10 for fine-tuning. Additionally, we reviewed key architectures frequently used in this field, including various families of Wide-ResNets, Vision Transformers, and handcrafted features combined with simple linear or convolutional layers.

In this Chapter, we focus on designing the datasets and benchmarks used in our work. As discussed in Chapter 3, a variety of architectures and datasets have been employed in the DP ML literature. To decide on the architectures and datasets, we considered several criteria, including which private datasets to use, which pre-training paradigms to adopt, what privacy accounting methods to employ, which parameters to select, and many more settings.

4.1 Datasets

In this section, we will describe the private datasets we use, and discuss the public datasets we rely on for pre-training our models. Additionally, we will explain why we chose these specific datasets and how they help us address the challenges of training DP ML models.

4.1.1 Private Datasets

For private datasets, we prescribe using the following medical datasets:

CheXpert. The CheXpert dataset [39] has 224,316 chest X-ray images of size 390×320 from 64,540 patients. Images may have multiple labels, where the possible labels correspond to five pathology classes: 'Cardiomegaly', 'Edema', 'Consolidation', 'Atelectasis', and 'Pleural Effusion'. In our work, following prior state-of-the-art training, we re-scale all images to size 224×224 and augment the dataset using random affine transformations [98, 99].

EyePACS. Kaggle EyePACS [36] contains retinal images of diverse populations with various degrees of diabetic retinopathy (DR). Each image is classified into one of five classes depending on the severity of the disease. The classification task is diagnostic of DR, as measured on a scale from 0 (no DR) to 4 (severe DR). The training set consists of 35,126 and the test set contains 53,576 color eye fundus images.

We chose medical images to address a privacy-critical setting where privacy is particularly relevant. Within this domain, chest X-rays and retinal fundus images are two of the most commonly used, so we selected one of the most popular datasets from each. Our selection was further guided by [74], which also focuses on medical image classification and examines CheXpert and a Google-proprietary diabetic retinopathy (DR) dataset. Furthermore, the study by Berrada et al. [7] also targets CheXpert to evaluate the efficacy of their private training approach on this dataset.

While there are several publicly available datasets for retinopathy detection, the majority are notably small, typically containing fewer than 100 images each. Such small datasets are not suitable for DP ML. Among them, EyePACS is the most popular one with an acceptable dataset size.

CheXpert and EyePACS datasets vary significantly across several dimensions, including the size of their training and testing sets, the nature of classification tasks (ranging from multi-class to multi-label), class imbalance issues, and they focus on very two different realms of medical imaging.

In addition to these two datasets, we recommend continuing to use well-established datasets like CIFAR-10 [52] and ImageNet [25] as standard benchmarks for training DP ML models from scratch, without using pre-training data. Although recent critiques have highlighted potential issues with these datasets [87], they are still widely used and allow for effective comparisons between private and non-private models.

4.1.2 Public Datasets

By exploring diverse pre-training regimes, we aim to understand the impact of different pre-training strategies on the performance and privacy of ML models. Each regime offers unique insights into how pre-training data influences the model's effectiveness in a privacy-preserving context. We consider the following datasets and regimes for pre-training:

ImageNet-1k pre-training: First, we recommend using ImageNet-1k as the public dataset for pre-training. This dataset is widely recognized as a large-scale image classification benchmark, making it a standard choice for pre-training in the field. Our approach is also inspired by the work of [24], who utilized this pre-training regime and achieved state-of-the-art performance on CIFAR-10 using Wide-ResNet architectures.

LAION-2B pre-training: Secondly, we suggest using LAION-2B, primarily because it is the pre-training data for OpenCLIP's ViT-G/14 model which is picked due to it's strong zero-shot capabilities. This dataset is substantial and offers a different distribution compared to ImageNet-1k, providing diverse pre-training data.

"Anything-goes" Regime: Lastly, we consider the general case of "anything-goes". In this regime, there are no limitations on the pre-training data, except for considerations to avoid data leakage where the public dataset contains private data. This track can incorporate large-scale Internet datasets, specialized domain-specific data, and other sources. Although the results from this category may not be directly comparable to one another, they offer a meaningful insight of overall progress on a benchmark. Additionally, we take into account the concerns with using public data for pre-training and propose the following regime.

No public data pre-training: Lastly, we consider the regime where we do not use any public data for pre-training. Instead, we directly perform private training on the private dataset. This can be achieved using handcrafted features that are never trained on images, thereby avoiding any additional privacy cost. Additionally, we study using synthetic data for pre-training as well. This approach helps create useful features without compromising privacy, as synthetic data significantly differs in its distribution from natural images.

4.2 Methods and Architectures

Differentially Private Stochastic Gradient Descent (DP-SGD) [79, 5, 1] is well-regarded for its versatility and model-agnostic nature. It can be used with various architectures without requiring modifications specific to each model. This broad applicability makes DP-SGD a popular choice for training ML models with DP.

In our study, we examine several key approaches to utilizing DP-SGD. One approach involves using logistic regression as a linear probe. This technique allows for efficient training and facilitates the comparison of various feature sets under DP constraints. By employing a linear probe, we can assess how different features contribute to model performance while maintaining privacy. Given the concerns regarding public data pre-training, it is crucial to create robust features that can compensate for the lack of pre-training. This approach is also particularly important when resources are limited and only last-layer fine-tuning is feasible. This regime is advantageous as it requires less computational power, making it suitable for environments with limited resources. We use a linear probe with handcrafted features [86], Wide-ResNet-16-4 [82], ResNet-28-10 [24], and CLIP ViT models [73].

Additionally, following [24], we also explore training from scratch and full-layer fine-tuning with ResNet-28-10 pre-trained on the ImageNet-1K dataset. This model is selected due to its proven performance in image classification tasks and inspired by the work of [24]. We also provide the PyTorch weights for this model. The Wide ResNet-28-10 serves as a robust and reliable choice for evaluating the impact of DP on a well-established image classification model.

4.2.1 Overview of Architectures and Techniques

Below is an overview of all the architectures and techniques we have employed in our study:

ScatterNets. Scattering Network (ScatterNet) [64] is a feature extractor that leverages wavelet transforms [13] to learn image representations that are robust to translation and small rotations. Tramèr and Boneh [86] applied these networks to MNIST, Fashion-MNIST, and CIFAR-10, using DP to train linear models and simple convolutional neural networks (CNNs) on top. Their findings show that using Scatter features with just a linear probe can outperform DP training of CNNs end-to-end on MNIST and Fashion-MNIST. For CIFAR-10, they demonstrated that a simple two-layer CNN trained on Scatter features yields better results than previous methods for DP training on CIFAR-10. We are particularly

interested in this architecture due to its strong performance in scenarios without public pre-training. Additionally, since the feature extractor is handcrafted and not trained on any dataset, it does not consume additional privacy budget, which is important in DP ML. For our experiments, we utilize ScatterNet to generate handcrafted features for both EyePACS and CheXpert, following the exact settings described in [86].

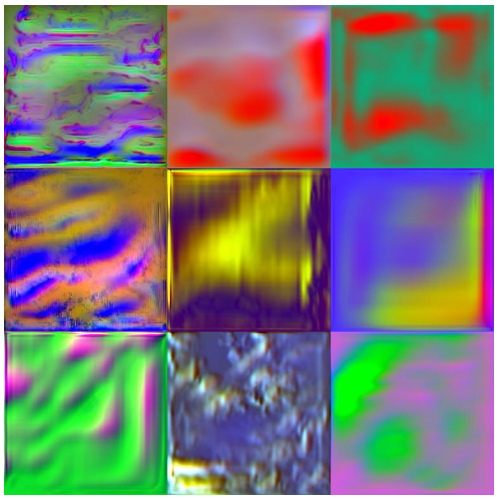


Figure 4.1: Synthetic data from random processes, StyleGAN [3].

Wide-ResNets. WideResNet [100] (WRN) stands out as a deep learning framework characterized by its wide yet shallow design, which improves feature learning and robustness in image classification by augmenting model width and employing residual connections. De et al. [24] utilize these models and achieve state-of-the-art (SOTA) DP performance across various scenarios on CIFAR-10. They explored both DP training from scratch and DP fine-tuning, after pre-training on ImageNet-1K [1]. They specifically used WRN-16-4 and WRN-40-4 for training from scratch and WRN-28-10 for fine-tuning. For our experiments, we use the WRN-28-10 architecture across all settings to ensure consistency. To facilitate direct comparison, we replicate their settings as closely as possible, incorporating techniques such as weight standardization [12], group normalization, and the same hyperparameters for pre-training. We conduct experiments both without any public data and with pre-training on ImageNet-1K.

Additionally, Tang et al. [82] utilize WRN-16-4 to achieve DP SOTA performance on CIFAR-10, when no extra public data is used for pretraining. They leverage image priors

[1]The ImageNet-1K data was downsampled to 32×32 pixels (referred to as IN-32 [21]).

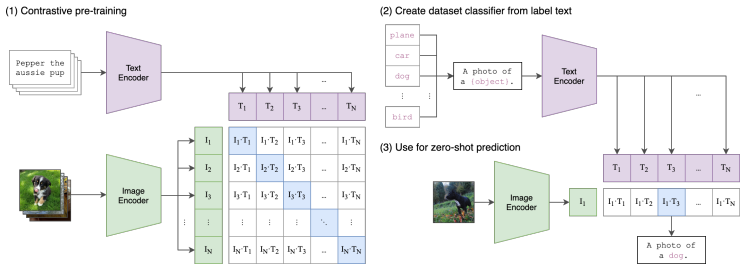


Figure 4.2: Overview of CLIP architecture from [73], which is a neural network trained on (image, text) pairs.

generated by random processes instead of starting from random initialization as shown in 4.1, outperforming [86] and [24] when they only train from scratch. Moreover, they achieve SOTA performance using only a linear probe, making for a direct comparison to the linear SN method of [86]. We adopt the same architecture and replicate their settings to the greatest extent possible, incorporating techniques such as augmentation multiplicity and normalization. Tang et al. [82] build on the approach of De et al. [3] by using the third-to-last layer of the network, which has a dimension of 4096. We adopt a similar strategy but reduce the dimensionality to 2048. This adjustment is necessary due to the larger image sizes in our datasets (CheXpert and EyePACS with 224×224 images) compared to CIFAR-10 (32×32 images) and resource constraints.

CLIP-based models. CLIP [73] is a popular contrastive learning pre-training technique, which allows one to jointly train a language and image encoder. CLIP has been observed to enable robust zero-shot image classification when pre-trained on very large Internet datasets. We use two ViT [27] models pre-trained using CLIP: OpenAI's ViT-B/16 (pre-trained on the proprietary WebImageText (WIT) dataset) and OpenCLIP's ViT-G/14 model (pre-trained on LAION-2B [76]).[2] Besides pre-training data, these models differ in their size (12 and 48 layers, respectively) and patch size (16 and 14, respectively). To run zero-shot experiments, we use these models in their original form. For DP fine-tuning experiments, we utilize only the image encoder as a feature extractor and then apply either a linear layer (i.e., logistic regression) or a two-layer neural network (TLNN) with tanh/tempered sigmoid activations [67] on top. We exclusively use CLIP-based models with their respective public pre-training datasets.

[2] https://github.com/mlfoundations/open_clip

4.3 Privacy Parameters

As shown in Section 3, previous studies have explored a variety of settings by reporting different values for ε and δ. This diversity in parameter makes it harder to effectively compare the results across different approaches. To facilitate a more straightforward and efficient comparison, we focus on a specific range of ε values: $\varepsilon = 1$, $\varepsilon = 3$, $\varepsilon = 5$, and $\varepsilon = 8$. These values span both lower and higher privacy levels, covering a broad spectrum of the typical range considered in DP ML. By using this range, we aim to provide a comprehensive overview of how different privacy levels impact the performance and effectiveness of various methods. Additionally, we suggest fixing the δ parameter to be to be the largest power of 10 that is at most the inverse of the training set size, similar to previous studies. While δ can often be adjusted with minimal impact on ε in many cases, this approach ensures consistency and comparability across different settings.

For every DP algorithm, there exists a privacy proof that provides an upper bound for ε and δ. In the case of DP-SGD algorithms, privacy accountants handle this process by taking various hyperparameters and δ as inputs to calculate the corresponding ε. Over time, advancements in privacy accounting have provided tighter bounds, with the development of "exact" privacy accounting methods [1, 60, 61, 51, 38]. These methods provide tighter bounds on ε and δ. However, recent studies [71, 22, 54] have shown that using a more refined accountant might give the false impression of a better result, even if the training process remains unchanged. To address this issue, it is important to report the privacy accounting method used, or the specific proof followed if not using DPSGD. We use PRV accountant in our experiments.

4.4 Leaderboard

Tracking progress on benchmark datasets via leaderboards is an established practice in (non-private) ML.[3] This is not yet the case for DP ML: a broad and up-to-date knowledge of the literature is required to keep track of the latest results, making entering the field especially challenging and intimidating for newcomers. As one of our contributions, we alleviate this issue by creating and maintaining a leaderboard for DP ML.[4]

Due to the particulars of the DP setting, it is unnatural to simply incorporate results into an existing leaderboard for the non-private setting. Specifically, beyond just the specific dataset, a leaderboard for DP ML would need to track many of the considerations

[3] See, e.g., https://paperswithcode.com/sota

[4] https://private-machinelearning.github.io/

already discussed, including the privacy parameters (ε, δ), which privacy accountant was used, and which public datasets were used. Another difference from the non-private setting is the issue of *correctness*. For a proposed algorithm, the DP guarantees must be mathematically *proven*, and a claimed result could be false if there is a bug in the proof. This is in addition to existing concerns from the non-private setting on whether results are independently reproducible or not. However, since it is notoriously easy to have bugs in a proof of DP, we incorporate a *verification* system to our leaderboard. By default, all results are unverified when added. However, anyone is able to submit a pull request to our GitHub to verify that they reproduced the result, and believe correctness of the privacy proof (if applicable).

At present, our leaderboard focuses exclusively on DP image classification (as does this paper), though it may be extended to other problems (e.g., DP natural language understanding or generation).

Chapter 5

Experiments and Results

In Chapter 4, we focused on dataset and benchmark design. We selected CheXpert and EyePACS as private datasets. We chose Wide-ResNets for their proven success in public data pre-training and fine-tuning. Vision Transformers (ViT) were included for their strong zero-shot capabilities and extensive pre-training on large datasets. We also looked into handcrafted features for scenarios with no public data pre-training. Inspired by Tang et al. [82], we also considered synthetic data pre-training with Wide-ResNets.

Beyond proposing a variety of datasets and evaluation settings for benchmarking, we experimentally investigate techniques and the resulting utility obtained therein. Some of the key questions guiding our exploration:

1. How many of the lessons learned in DP image classification on datasets like CIFAR-10 and ImageNet transfer to the privacy-critical setting of medical images?
2. How much and when does public data help for such datasets, which may be ill-represented in public data?
3. In absolute terms, how well can we do on these datasets with DP, in various evaluation settings?

After describing our experimental setup in Section 5.1, we revisit the efficacy of several ablations commonly employed in DP settings in Section 5.2. Finally, we make more broad conclusions about DP image classification based on our results in Section 5.3.

5.1 Experimental Setup

We use PyTorch [69], and the Opacus library [93] for DP ML. We employed the Adam optimizer [49] across all experiments, both private and non-private, with a default learning rate of 0.001. We run our experiments at a variety of privacy levels ($\varepsilon \in [1, 3, 5, 8]$) with fixed delta values proportional to the inverse of the dataset size (10^{-6} for CheXpert and 10^{-5} for EyePACS), as we prescribed earlier. Batch size and total training epochs were fixed at 1024 and 20, respectively. A hyper-parameter search was performed to identify the optimal clipping norm within the range $[0.001, 0.01, 0.1, 1, 10]$. Following established metrics for all these datasets, we use AUC for CheXpert and EyePACS, and accuracy for CIFAR-10. We report mean and standard deviation over three independent runs.[1]

We used early stopping for non-private numbers due to overfitting, a phenomenon we did not observe for the DP setting due to its natural regularization properties [45]. Except for the ablation study, which examines the effects of various normalizations, data normalization is always applied for EyePACS Scatter features to ensure the model converges.

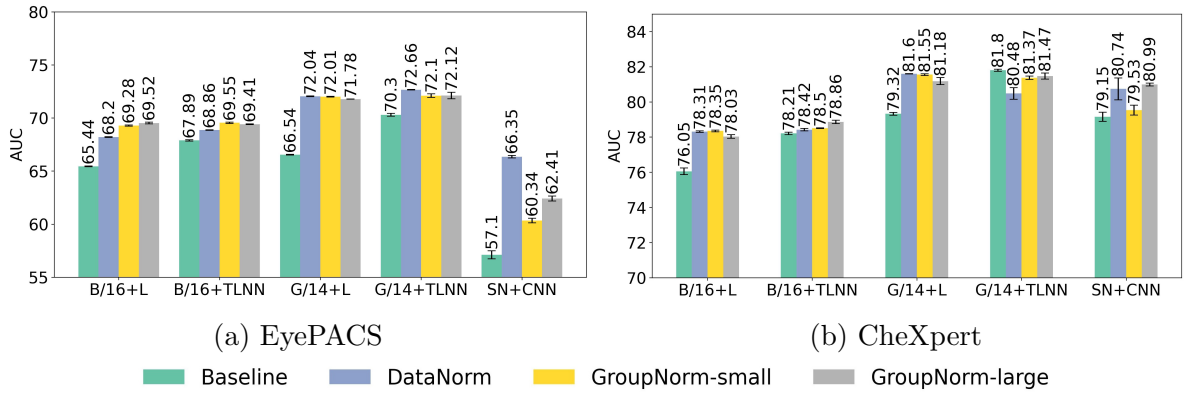
5.2 Revisiting DP ablations

One of the most comprehensive ablation studies for DP image classification is by De et al. [24]. By using group normalization, large batch sizes, weight standardization, augmentation multiplicity, and parameter averaging, they manage to raise CIFAR-10 accuracy on a validation set from 50.8% to an impressive 79.7%. We fix $\varepsilon = 3$ and, focusing on the CLIP ViTs and ScatterNets, run the exact same ordered sequence of ablations, without carrying forward the latest technique if it does not show improved utility. Broadly speaking, while De et al. [24]'s techniques proved highly effective for CIFAR-10, our results reveal mixed outcomes depending on various parameters.

Normalization. Batch normalization [43] is not compatible with DPSGD because it combines information across a batch, making it impossible to bound the impact of a single image in the dataset. Instead, prior work has shown that variants including group normalization [90] and data normalization [86] can be suitable replacements [86, 24, 7, 34].

Group normalization splits the channels of the hidden activations of an image into groups and normalizes them within each group. For the CLIP ViT models, with an input

[1] Some exceptions include training WRN models from scratch, due to compute limitations.



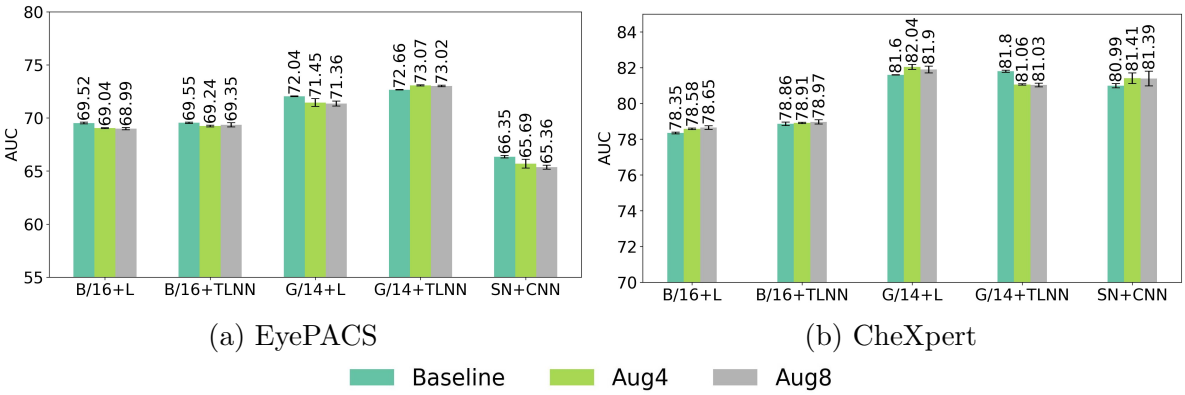
(a) EyePACS (b) CheXpert

Figure 5.1: Normalization generally improves the final performance of all models. The yellow bars correspond to groups of 8 and 9, while the gray bars correspond to groups of 16 and 27 for CLIP-ViTs and ScatterNet, respectively. The choice of 27 over 81 is due to its superior performance over 9.

dimension of 512 and 1280 for the B/16 and G/14 respectively, we experiment with 8 and 16 groups. For Scatter features, with dimension (243, H/4, W/4) for RGB images and following Tramèr and Boneh [86], we use 9, 27, and 81 groups. Data normalization works on data channels by normalizing using the corresponding mean and variance across the training data. Normalizing in such a way, however, incurs a privacy cost as the per-channel means and variances must be privately estimated. We use Gaussian noise with ($\sigma = 8$) to estimate these means and variances for all runs, following Tramèr and Boneh [86].

In Figure 5.1 we show that normalization generally improves the final performance of all models, though the most effective normalization differs across architectures and datasets. Interestingly, the four experiments where data normalization was superior involved models with larger unclipped gradients. In these cases, the optimal clipping norm chosen during hyperparameter tuning was also the highest value, i.e., 10. This suggests that data normalization can effectively manage large gradient magnitudes, especially when clipping underestimates the true gradient norms.

Larger Batch Size. The impact of larger batch sizes in differentially private training has been observed both theoretically [4, 81] and empirically [2, 24]. In Table A.4, scaling the batch size from 1024 to 4096 showed that CheXpert benefited in 80% of experiments, while EyePACS did not. This disparity is likely due to CheXpert having a training set six times larger than EyePACS, resulting in fewer model update steps for EyePACS and potential underfitting with a fixed number of epochs. We further observed that increasing



(a) EyePACS (b) CheXpert

Figure 5.2: Evaluating augmentation multiplicity by adding 4 and 8 augmentations of each image in the training data. Augmentation multiplicity helps in general for CheXpert but not for EyePACS.

the number of epochs showed a positive impact of larger batch sizes on EyePACS when using the ScatterNet model which supported our hypothesis.

Weight Standardization We experiment with weight standardization (WS) only on the Scatternet + CNN model since it applies to only convolution layers. From our results in Table A.4, we observe that WS does not help with EyePACS but helps with CheXpert and CIFAR-10. As alluded by prior work [12, 24], we also observe a positive correlation of group normalization with WS. However, due to a limited number of experiments, we do not have strong evidence either way.

Augmentation Multiplicity. We apply a sequence of augmentations to our benchmark datasets: reflect padding, random cropping, and random horizontal flipping. While De et al. [24] recommend 16 augmentations per image, due to computational constraints with large datasets, we use 4 and 8 augmentations. As shown in Figure 5.2, contrary to De et al. [24]'s findings, augmentation multiplicity (augmult) does not consistently yield positive effects. Except for one experiment, (ViT-G/14+TLNN), augmentations generally benefit CheXpert but not EyePACS. Future work may explore the effectiveness of dataset-specific augmentations, which could potentially yield more beneficial results.

Parameter Averaging The final ablation that De et al. [24] suggests is the exponential moving average (EMA) [70] of all the parameters in the model. In Table A.4, we notice that

EMA occasionally improves performance, which contradicts the findings of De et al. [24] that it consistently enhances results across all experiments.

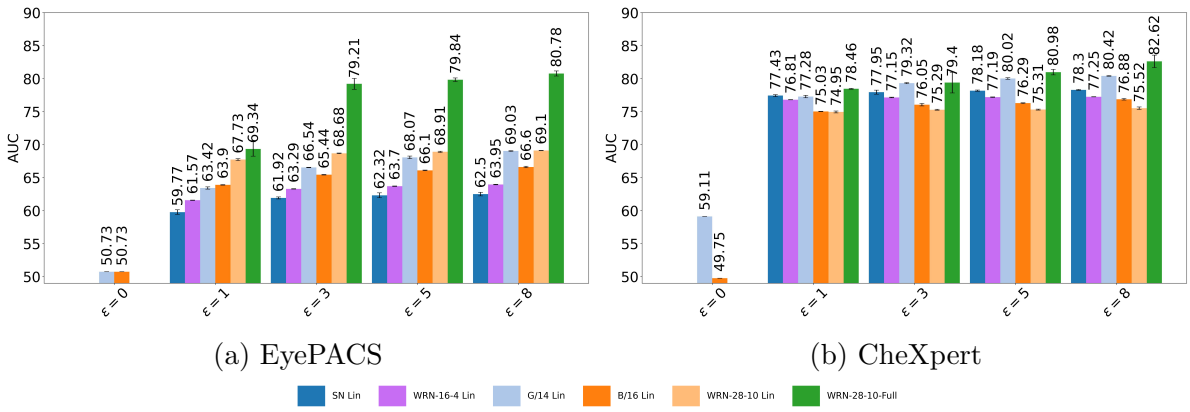
5.3 Experimental Findings

We highlight some findings from our experimental results.

Different pre-training datasets and architectures offer varying degrees of improvement depending on the private data. We compare models publicly pre-trained on a variety of datasets on both CheXpert and EyePACS. Results are displayed in Figure 5.3. For the case of no pre-training data, we choose ScatterNet Linear and Wide-ResNet16-4 Linear, due to their consistently superior utility compared to Wide-ResNet-28-10 trained from scratch, particularly for high privacy (i.e., low ε) settings. On the other end of the spectrum, when we allow large-scale public pre-trarining, the CLIP ViT models provide a good indication of zero-shot performance (i.e., $\varepsilon = 0$).

Wide-ResNet pre-trained on ImageNet provides the most benefit for EyePACS, whereas ViT-G/14 pre-trained on LAION-2B is more effective for CheXpert; For EyePACS, when comparing Wide-ResNet models pre-trained on ImageNet, we observe a significant improvement going from training from scratch to full fine-tuning, with the performance increasing from 55.51% to 69.34% at $\varepsilon = 1$. Interestingly, this gap widens as ε increases. In contrast, for CheXpert, the improvement from training from scratch to full fine-tuning is more modest, with the AUC increasing from 76.87% to 78.46% at $\varepsilon = 1$. The non-private results for CheXpert show a smaller difference, with the AUC close at 87.28% for training from scratch and 87.57% for full fine-tuning, whereas, for EyePACS we observe that training from scratch achieves 62.15% and full fine-tuning results in 83.66% AUC. Additionally, we observe that transitioning from scratch to linear fine-tuning adversely affects performance on CheXpert, while the opposite effect is observed for EyePACS.

When analyzing CheXpert, ViT-B/16 performs close to random in the zero-shot setting, whereas ViT-G/14 achieves a nontrivial AUC of 59.11%. Moving from ScatterNet+linear to Wide-ResNet+Linear, there is a noticeable decrease in AUC, yet ViT-G/14 consistently outperforms across various ε values, indicating that ViT-G/14 pre-trained on LAION is a better fit for CheXpert. Notably, at $\varepsilon = 8$, Wide-ResNet with full fine-tuning exceeds the performance of ViT-G/14. However, considering that Wide-ResNet is fully fine-tuned while ViT-G/14 is not, this doesn't necessarily make Wide-ResNet better suited for CheXpert.



(a) EyePACS (b) CheXpert

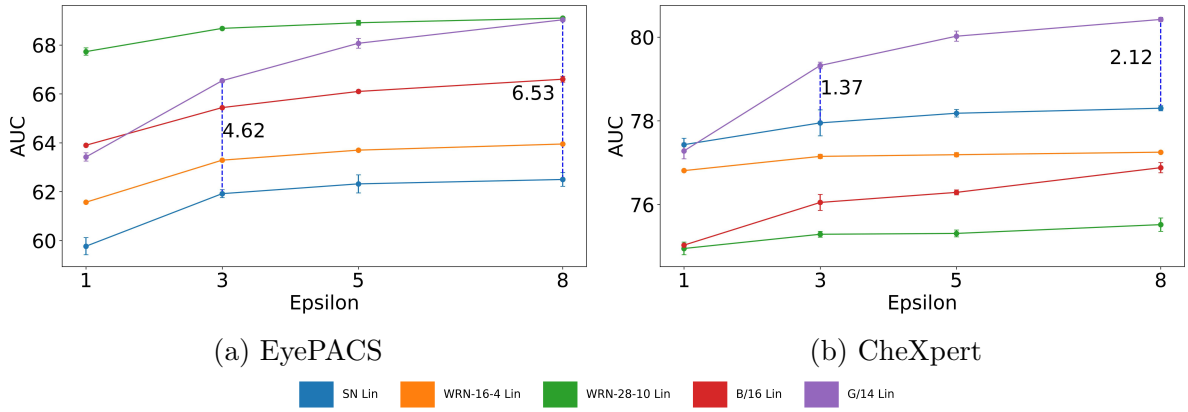
Figure 5.3: Pre-training datasets have varying impacts: Wide-ResNet pre-trained on ImageNet benefits EyePACS most, while ViT-G/14 with linear zero-shot significantly improves CheXpert performance from random guessing, outperforming Wide-ResNet in linear fine-tuning across different ε values and approaching fully fine-tuned Wide-ResNet performance.

Nevertheless, for smaller ε, it is clear that ViT-G/14 is the superior model with only linear fine-tuning.

Looking at Figure 5.3 for EyePACS, both CLIP ViT models show random performance in the zero-shot setting, indicating no improvement from pre-training. Conversely, Wide-ResNet linear exhibits a significant performance boost when transitioning from ScatterNet linear to Wide-ResNet linear, maintaining its superiority across all ε values. Although we notice that as we move toward less private regimes, the power of pre-trained ViT-G/14 becomes more evident, particularly from $\varepsilon = 1$ to $\varepsilon = 3$, approaching the performance of Wide-ResNet linear. However, there remains a substantial gap between fully fine-tuned Wide-ResNet and the other models, unlike CheXpert, suggesting that Wide-ResNet pre-trained on ImageNet is better suited for EyePACS than it is for CheXpert.

Public pre-training data helps more with higher ε values. In Figure 5.4, comparing linear probes on CheXpert, ScatterNet achieves the best result at $\varepsilon = 1$. However, as ε increases, pre-trained models, especially ViT-G/14, begin to outperform other methods significantly. While full fine-tuning of CLIP has not been explored, a direct comparison of features shows ViT-G/14's superiority when ε is sufficiently large. As ε value increases further, ViT-G/14's performance improves notably, highlighting its strong pre-trained performance under less stringent privacy constraints.

As shown in Figure 5.4, we see that for ViT-G/14, increasing the ε value widens the



(a) EyePACS (b) CheXpert

Figure 5.4: Pre-training public data is more beneficial with higher ε values. For CheXpert, ScatterNet performs better at smaller ε values, while pre-trained models show marginal improvements at larger ε values. Similarly, for EyePACS, CLIP ViT-G/14 linear performs better as ε value increases.

gap between it and the ScatterNet baseline. However, we do not observe any significant changes in ViT-B/16 and Wide-ResNet linear. ViT-B/16 appears to perform poorly regardless of privacy settings. On the other hand, Wide-ResNet linear consistently maintains a significant gap between it's linear model and ScatterNet. This can be explained by the fact that Wide-ResNet linear can already achieve high AUC in the $\varepsilon = 1$ case, leaving little room for improvement.

The fact that Wide-ResNet maintains its advantage from the start is not surprising. As noted, the pre-trained Wide-ResNet model has shown significant benefits for the EyePACS dataset, likely due to its robustness and the advantages gained from prior training on public data. On the other hand, the performance of the ViT-G/14 model exhibits a different trend. As the ε value increases—indicating a relaxation of privacy constraints—the model's performance notably improves. This trend suggests that the ViT-G/14 model benefits more from the increased privacy budget.

Progress on CIFAR10 does not translate to progress on benchmark datasets. Upon reviewing Figure 5.3, it becomes clear that the performance of the ViT-G/14 model varies significantly across different datasets. On CIFAR-10, ViT-G/14 achieves an impressive 99.75% zero-shot accuracy, highlighting its exceptional capability to generalize on this dataset. However, when applied to CheXpert and EyePACS, the model's performance deteriorates sharply, with AUC scores of 59.11% and 50.73%, respectively. The score on

EyePACS is notably low, effectively equating to random guessing, which underscores the model's limited effectiveness in this context.

Similarly, the Wide-ResNet model demonstrates strong performance on CIFAR-10, achieving 94.7% accuracy with a privacy parameter of $\varepsilon = 1$. Yet, its performance on CheXpert and EyePACS is significantly lower compared to non-private numbers it can achieve, with AUC scores of 78.52% and 71.00%, respectively. This stark contrast illustrates the challenge of transferring model performance from one domain to another, particularly from CIFAR-10 to qualitatively different domains, i.e., medical datasets like CheXpert and EyePACS.

Additionally, we observe that incorporating synthetic data for pretraining, as demonstrated by Tang et al. [82], leads to state-of-the-art (SOTA) performance on CIFAR-10 without pre-training. However, in our experiments, ScatterNet outperforms [82]'s approach on CheXpert, whereas on EyePACS, [82] achieve better results.

Upon reviewing the ablation experiments presented in Section 5.2, it is clear that the techniques that proved beneficial for CIFAR-10 do not necessarily offer the same advantages for the EyePACS and CheXpert datasets. The performance patterns observed with CIFAR-10 did not consistently translate to these medical image datasets. In particular, we found significant variations in performance across different techniques when applied to CheXpert and EyePACS, highlighting inconsistencies in their effectiveness. This discrepancy underscores the need for dataset-specific strategies and adjustments, as the characteristics and challenges of medical image datasets differ markedly from those of standard image classification benchmarks.

[2]The authors use an ensemble of five models.

Table 5.1: Test AUC evaluation for EyePACS at different privacy regimes. Baselines include ScatterNet (SN), WideResNet-28-10 (WRN), WideResNet-16-4 (WRN-16-4), and CLIP models on datasets with different public data pre-training. The SOTA is due to [88].

Public	Model	Test AUC (%)				
		$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 5$	$\varepsilon = 8$	$\varepsilon = \infty$
None	SN + L	59.77 ± 0.35	61.92 ± 0.16	62.32 ± 0.37	62.5 ± 0.28	69.70 ± 0.11
None	SN + CNN	63.73 ± 0.11	66.36 ± 0.17	66.59 ± 0.43	67.37 ± 0.27	69.28 ± 0.20
None	WRN-16-4+L	61.57 ± 0.02	63.29 ± 0.04	63.7 ± 0.04	63.95 ± 0.04	67.74 ± 0.01
None	WRN (Scratch)	55.45 ± 0.18	56.53 ± 0.08	57.14 ± 0.34	57.65 ± 0.22	61.97 ± 0.09
IN-32	WRN+L	67.73 ± 0.15	68.68 ± 0.05	68.91 ± 0.10	69.10 ± 0.03	73.21 ± 0.09
IN-32	WRN (Full)	69.34 ± 1.09	79.21 ± 0.83	79.84 ± 0.27	80.78 ± 0.38	83.61 ± 0.03
WIT	B/16+L	63.9 ± 0.04	65.44 ± 0.04	66.1 ± 0.05	66.6 ± 0.12	69.93 ± 0.01
WIT	B/16+TLNN	65.12 ± 0.04	67.89 ± 0.06	69.22 ± 0.1	69.84 ± 0.02	70.54 ± 0.01
LAION	G/14+L	63.42 ± 0.17	66.54 ± 0.04	68.07 ± 0.2	69.03 ± 0.06	69.88 ± 0.2
LAION	G/14+TLNN	65.47 ± 0.02	70.30 ± 0.13	71.74 ± 0.3	72.3 ± 0.19	73.36 ± 0.15
IN-1K	SOTA	-	-	-	-	95.1

Table 5.2: Test AUC evaluation for CheXpert at different privacy regimes. Baselines include ScatterNet (SN), WideResNet-28-10 (WRN), WideResNet-16-4 (WRN-16-4), and CLIP models on datasets with different public data pre-training. The SOTA is due to [7](Private) and [98](Non-Private).

Public	Model	Test AUC (%)				
		$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 5$	$\varepsilon = 8$	$\varepsilon = \infty$
None	SN+L	77.43 ± 0.15	77.95 ± 0.31	78.18 ± 0.09	78.30 ± 0.06	78.94 ± 0.13
None	SN+CNN	78.16 ± 0.22	79.15 ± 0.26	79.16 ± 0.18	79.68 ± 0.04	80.65 ± 0.12
None	WRN-16-4+L	76.81 ± 0.02	77.15 ± 0.05	77.19 ± 0.05	77.25 ± 0.01	77.49 ± 0.03
None	WRN (Scratch)	76.9 ± 0.02	77.8 ± 0.04	77.89 ± 0.05	78.68 ± 0.10	87.31 ± 0.07
IN-32	WRN+L	74.95 ± 0.15	75.29 ± 0.07	75.31 ± 0.08	75.52 ± 0.16	75.91 ± 0.08
IN-32	WRN (Full)	78.46 ± 0.07	79.40 ± 1.57	80.98 ± 0.42	82.62 ± 0.94	87.62 ± 0.09
WIT	B/16+L	75.03 ± 0.03	76.05 ± 0.19	76.29 ± 0.06	76.88 ± 0.12	76.89 ± 0.01
WIT	B/16+TLNN	77.28 ± 0.19	78.21 ± 0.07	78.33 ± 0.04	78.54 ± 0.06	78.56 ± 0.01
LAION	G/14+L	77.28 ± 0.19	79.32 ± 0.08	80.02 ± 0.12	80.42 ± 0.05	80.48 ± 0.02
LAION	G/14+TLNN	80.63 ± 0.4	81.80 ± 0.06	82.25 ± 0.02	$82.27 \pm 0.0.4$	82.28 ± 0.01
IN-21K	SOTA	86.3	-	-	89.2	-
IN-1K	SOTA	-	-	-	-	93 [2]

Chapter 6

Conclusion and Future Work

In this Chapter, we summarize the thesis and the key contributions in Section 6.1, and discuss future work in Section 6.2.

6.1 Summary of Thesis and Contributions

The main motivation of this thesis stems from some important critiques raised by Tramèr, Kamath, and Carlini [87] regarding several common trends in differentially private machine learning (DP ML). Firstly, much of the research has focused on datasets that may not accurately represent privacy-sensitive domains. For example, many studies use datasets like CIFAR-10, which contains natural images, while less attention has been given to more privacy-critical areas, such as medical datasets. Additionally, although there has been extensive work on pre-training methods, these studies often involve scenarios where the public and private data are qualitatively similar. This raises questions about whether these methods truly enhance private training or it only performsn intra-domain transfer. Furthermore, the use of public datasets poses its own challenges. These datasets may contain sensitive information and are often curated with minimal oversight, raising concerns about their appropriateness for privacy-sensitive applications. We discussed this problem statement in detail in Chapter 2.

Our hypothesis was that advancements made with widely used datasets like CIFAR-10 might not apply as effectively to more privacy-sensitive domains, such as medical datasets. Since these medical datasets are qualitatively different from common public datasets like ImageNet, which are often used for pre-training in DP ML, we anticipated that achieving

similar results through pre-training would be more challenging. Our goal was to evaluate the concerns raised by Tramèr, Kamath, and Carlini [87] and provide insights into how progress in DP ML can be better assessed and applied to privacy-sensitive datasets.

In Chapter 4, we proposed standardized benchmark datasets and evaluation settings to assess progress in DP image classification, focusing particularly on privacy-sensitive domains using two prominent medical datasets: CheXpert and EyePACS. We evaluated previously established techniques for private and non-private image classification across various settings to determine their effectiveness. Specifically, we investigated architectures such as Wide-ResNets, CLIP ViT models, and Scattering Networks, and examined a range of training approaches, including training from scratch, full fine-tuning, last-layer fine-tuning (linear probe), and shallow CNN architectures on handcrafted features. These approaches have previously demonstrated strong performance on well-known datasets like CIFAR-10 and ImageNet.

In Chapter 5, we conducted extensive experiments based on the benchmarks described in Chapter 4 and evaluated the results. These findings supported our initial hypothesis. Here, we briefly outline the key findings:

- **Transferability of Techniques:** Our experiments revealed that improvements observed on CIFAR-10 did not necessarily transfer to the medical datasets we examined. Many established best practices for enhancing DP ML performance on CIFAR-10, such as augmentation multiplicity and weight standardization, were less effective on our medical datasets and, in some cases, even detrimental. These techniques appeared to be highly dependent on the specific data and architecture used. However, normalization generally proved beneficial across experiments. The extent of its usefulness varied depending on the context; for example, with handcrafted features for EyePACS, where gradients were large, data normalization was crucial for convergence. In contrast, group normalization demonstrated better performance in other scenarios. Therefore, the advancements achieved on CIFAR-10 do not necessarily translate to similar improvements on other datasets.

- **Effectiveness of Pre-Training:** Pre-training remained valuable, especially in full fine-tuning scenarios, but its effectiveness varied by dataset. For instance, Wide-ResNet pre-trained on ImageNet was most beneficial for EyePACS, while ViT-G/14 with two-layer linear probing outperformed Wide-ResNet with full fine-tuning on CheXpert across various ε values, particularly in tighter privacy regimes. Additionally, ViT-G/14 zero-shot significantly improved performance on CheXpert compared to random guessing, although this did not apply to EyePACS. These findings high-

light that different pre-training datasets provide varying levels of improvement depending on the private data. This is particularly important for privacy-sensitive domains, which are often underrepresented in public datasets.

- **Performance in Tighter Privacy Regimes:** In tighter privacy regimes (lower ε values, particularly when $\varepsilon = 1$), we observed that handcrafted features outperformed pre-trained models for linear probing on CheXpert. This indicates that while pre-trained models are beneficial, handcrafted features can still offer advantages in more restrictive privacy settings which is a more desirable regime. This suggests potential for further advancements in DP ML, especially in cases where public data pre-training is limited, especially due to privacy concerns, as detailed in Chapter 2. Handcrafted features and Synthetic data pre-training continue to hold significant value in tighter privacy regimes.

- **Public Leaderboard:** To support the community, we released a public leaderboard for DP ML, allowing for tracking progress on these benchmarks and fostering further research and development in this area.

6.2 Future Work

6.2.1 More Datasets and Architectures

In this work, we mainly focused on two well-known medical datasets. However, there is a wide range of other datasets that also deserve attention in privacy domain. They can cover different areas beyond just chest X-rays and retinopathy. There are definitely more diverse benchmarks that could be used to assess performance on the introduced datasets. These include, but are not limited to, other ResNet families and Vision Transformers (ViTs). Additionally, exploring more training approaches, such as training CLIP models from scratch (if resources permit), could be particularly interesting given CLIP's strong zero-shot capabilities, especially with CheXpert. However, this could be challenging because adapting the architecture to be compatible with DP involves replacing batch normalization layers, which may necessitate training from scratch since pre-trained weights can't be transferred anymore. Despite these challenges, if sufficient resources are available, pursuing this could provide valuable insights.

6.2.2 Language Models

The scope for further research is extensive, beyond image classification, such as language models. In language processing, pre-training plays a crucial role in developing models with strong performance. As large language models (LLMs) become more prevalent and their capabilities expand, it is increasingly important to address privacy concerns related to the pre-training data used. Much of the data for these models is scraped from the internet, often with minimal oversight, which raises significant privacy issues. To mitigate these concerns, creating synthetic data has become an increasingly important approach. By using synthetic data for pre-training, one can reduce the risk of inadvertently exposing private information. This approach is especially pertinent in the language domain, where large-scale data is crucial for training effective models. As the need for privacy-preserving techniques grows, developing DP methods will become a key focus.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, 2016. ACM.

[2] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. *arXiv preprint arXiv:2108.01624*, 2021.

[3] Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2556–2569. Curran Associates, Inc., 2021.

[4] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[5] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society.

[6] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

[7] Leonard Berrada, Soham De, Judy Hanwen Shen, Jamie Hayes, Robert Stanforth, David Stutz, Pushmeet Kohli, Samuel L Smith, and Borja Balle. Unlocking ac-

curacy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.

[8] Alex Bie, Gautam Kamath, and Guojun Zhang. Private GANs, revisited. *Transactions on Machine Learning Research*, 2023.

[9] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision*, WACV '21, pages 1536–1546. IEEE, 2021.

[10] Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *CoRR*, abs/2101.08692, 2021.

[11] Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *CoRR*, abs/2101.08692, 2021.

[12] Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *CoRR*, abs/2101.08692, 2021.

[13] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35:1872–1886, 08 2013.

[14] G. J. Burton and Ian R. Moorhead. Color and spatial structure in natural scenes. *Appl. Opt.*, 26(1):157–170, Jan 1987.

[15] Tianshi Cao, Alex Bie, Arash Vahdat, Sanja Fidler, and Karsten Kreis. Don't generate me: Training differentially private generative models with sinkhorn divergence. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 12480–12492. Curran Associates, Inc., 2021.

[16] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, USENIX Security '21, pages 2633–2650. USENIX Association, 2021.

[17] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, USENIX Security '23, pages 5253–5270. USENIX Association, 2023.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[19] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020.

[20] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[21] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the CIFAR datasets. *CoRR*, abs/1707.08819, 2017.

[22] Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. How private are DP-SGD implementations? In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR, Inc., 2024.

[23] Anindya De, Philip M. Long, and Rocco A. Servedio. Learning sums of independent random variables with sparse collective support. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '18, pages 297–308, Washington, DC, USA, 2018. IEEE Computer Society.

[24] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 248–255, Washington, DC, USA, 2009. IEEE Computer Society.

[26] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

[28] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[29] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, EUROCRYPT '06, pages 486–503, Berlin, Heidelberg, 2006. Springer.

[30] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[31] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[32] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

[33] Friedrich Dörmann, Osvald Frisk, Lars Nørvang Andersen, and Christian Fischer Pedersen. Not all noise is accounted equally: How differentially private learning benefits from large sampling rates, 2021.

[34] Friedrich Dörmann, Osvald Frisk, Lars Nørvang Andersen, and Christian Fischer Pedersen. Not all noise is accounted equally: How differentially private learning benefits from large sampling rates, 2021.

[35] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

[36] Eyepacs. Diabetic retinopathy detection, 2015. data retrieved from Kaggle, https://www.kaggle.com/c/diabetic-retinopathy-detection.

[37] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the*

2015 ACM Conference on Computer and Communications Security, CCS '15, pages 1322–1333. ACM, 2015.

[38] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 11631–11642. Curran Associates, Inc., 2021.

[39] Stanford ML Group. Chexpert: A large chest x-ray dataset and competition. https://stanfordmlgroup.github.io/competitions/chexpert/. Accessed: 2024-06-01.

[40] Frederik Harder, Milad Jalali, Danica J Sutherland, and Mijung Park. Pre-trained perceptual features improve differentially private image generation. *Transactions on Machine Learning Research*, 2023.

[41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019.

[42] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

[43] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML '15, pages 448–456. JMLR, Inc., 2015.

[44] Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 12 2019.

[45] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *Proceedings of the 11th Conference on Innovations in Theoretical Computer Science*, ITCS '20, pages 31:1–31:17, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[46] Gautam Kamath. Cs 860: Algorithms for private data analysis fall 2020, lecture 5 — approximate differential privacy. http://www.gautamkamath.com/CS860notes/lec5.pdf. Accessed: 2024-08-08.

[47] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *CoRR*, abs/1912.04958, 2019.

[48] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. *CoRR*, abs/2101.08515, 2021.

[49] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, 2015.

[50] Helena Klause, Alexander Ziller, Daniel Rueckert, Kerstin Hammernik, and Georgios Kaissis. Differentially private training of residual networks with scale normalisation, 2022.

[51] Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using FFT. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, AISTATS '20, pages 2560–2569. JMLR, Inc., 2020.

[52] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[53] Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *CoRR*, abs/2201.12328, 2022.

[54] Christian Janos Lebeda, Matthew Regehr, Gautam Kamath, and Thomas Steinke. Avoiding pitfalls for privacy accounting of subsampled mechanisms under composition. *arXiv preprint arXiv:2405.20769*, 2024.

[55] Yann LeCun, Corinna Cortes, and Chris Burges. MNIST handwritten digit database, 2010.

[56] Ann B. Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *Int. J. Comput. Vision*, 41(1–2):35–59, jan 2001.

[57] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

[58] Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

[59] Ilya Mironov. Renyi differential privacy. *CoRR*, abs/1702.07476, 2017.

[60] Ilya Mironov. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, CSF '17, pages 263–275, Washington, DC, USA, 2017. IEEE Computer Society.

[61] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[62] Milad Nasr, Reza Shokri, and Amir houmansadr. Improving deep learning with differential privacy using gradient encoding and denoising, 2020.

[63] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. *CoRR*, abs/1412.8659, 2014.

[64] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. *CoRR*, abs/1412.8659, 2014.

[65] Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. https://openreview.net/forum?id=rJg851rYwH, 2019.

[66] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate, 2018.

[67] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 9312–9321, 2021.

[68] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy, 2020.

[69] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 8026–8037. Curran Associates, Inc., 2019.

[70] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 07 1992.

[71] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

[72] Katyanna Quach. Inside the 1tb imagenet data set used to train the world's ai: Naked kids, drunken frat parties, porno stars, and more. https://www.theregister.com/2019/10/23/ai_dataset_imagenet_consent. Accessed: 2024-07-04.

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[74] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging, 2019.

[75] Tom Sander, Pierre Stock, and Alexandre Sablayrolles. Tan without a burn: Scaling laws of dp-sgd, 2023.

[76] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[77] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE*

Symposium on Security and Privacy, SP '17, pages 3–18, Washington, DC, USA, 2017. IEEE Computer Society.

[78] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the 2023 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '23, pages 6048–6058. IEEE Computer Society, 2023.

[79] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing*, GlobalSIP '13, pages 245–248, Washington, DC, USA, 2013. IEEE Computer Society.

[80] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *CoRR*, abs/1707.02968, 2017.

[81] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

[82] Xinyu Tang, Ashwinee Panda, Vikash Sehwag, and Prateek Mittal. Differentially private image classification by learning priors from random processes, 2023.

[83] Xinyu Tang, Ashwinee Panda, Vikash Sehwag, and Prateek Mittal. Differentially private image classification by learning priors from random processes, 2023.

[84] David Thiel. Identifying and eliminating csam in generative ml training data and models. https://purl.stanford.edu/kh752sm9123. Accessed: 2024-07-18.

[85] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[86] Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021.

[87] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR, Inc., 2024.

[88] Mike Voets, Kajsa Møllersen, and Lars Ailo Bongo. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLOS ONE*, 14(6):e0217541, June 2019.

[89] Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David O'Brien, Thomas Steinke, and Salil Vadhan. Differential privacy: A primer for a non-technical audience. *SSRN Electronic Journal*, 01 2018.

[90] Yuxin Wu and Kaiming He. Group normalization. *CoRR*, abs/1803.08494, 2018.

[91] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[92] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

[93] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Gosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[94] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

[95] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021.

[96] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. *CoRR*, abs/2106.09352, 2021.

[97] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2019.

[98] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

[99] Zhuoning Yuan, Dixian Zhu, Zi-Hao Qiu, Gang Li, Xuanhui Wang, and Tianbao Yang. Libauc: A deep learning library for x-risk optimization. In *29th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[100] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[101] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

APPENDICES

Appendix A

Ablations Study Plots

In this chapter, we present detailed tables from the ablation study discussed in Chapter 5. These tables include results for various normalization techniques, augmentation multiplicity, and other ablations.

A.1 Normalization

A.1.1 CLIP ViT Architectures

Table A.1: Studying the impact of normalization for CLIP ViT models, we find that normalization generally improves performance, but the effectiveness depends on the architecture and dataset. Normalizations marked in red indicate a drop in performance compared to the baseline.

Dataset	Model	Baseline	DataNorm	GroupNorm8	GroupNorm16
EyePACS	B/16 + Linear	65.44 ± 0.04	68.2 ± 0.04	69.28 ± 0.06	$\mathbf{69.52 \pm 0.08}$
EyePACS	B/16 + TLNN	67.89 ± 0.06	68.86 ± 0.03	$\mathbf{69.55 \pm 0.06}$	69.41 ± 0.04
EyePACS	G/14 + Linear	66.54 ± 0.04	$\mathbf{72.04 \pm 0.03}$	72.01 ± 0.04	71.78 ± 0.01
EyePACS	G/14 + TLNN	70.30 ± 0.13	$\mathbf{72.66 \pm 0.03}$	72.1 ± 0.17	72.12 ± 0.31
CheX	B/16 + Linear	76.05 ± 0.19	78.31 ± 0.04	$\mathbf{78.35 \pm 0.05}$	78.03 ± 0.1
CheXpert	B/16 + TLNN	78.21 ± 0.07	78.42 ± 0.07	78.5 ± 0.02	$\mathbf{78.86 \pm 0.1}$
CheXpert	G/14 + Linear	79.32 ± 0.08	$\mathbf{81.6 \pm 0.01}$	81.55 ± 0.05	81.18 ± 0.2
CheXpert	G/14 + TLNN	$\mathbf{81.80 \pm 0.06}$	80.48 ± 0.33	81.37 ± 0.1	81.47 ± 0.17
CIFAR10	CLIP + Linear	99.64(93.91)	99.76(94.41)	99.75(94.43)	**99.75(94.57)**
CIFAR10	CLIP + TLNN	99.69(94.10)	99.74(94.15)	99.74(94.36)	**99.75(94.51)**

A.1.2 ScatterNets

Table A.2: Studying the impact of normalization for ScatterNet + CNN, we observe that normalization consistently improves performance. Data normalization tends to outperform group normalization for EyePACS and CIFAR-10, particularly due to the large gradients of their Scatter features.

Dataset	Model	Baseline	DataNorm	GN 9	GN 27	GN 81
EyePACS(AUC)	SN + CNN	57.1 ± 0.38	$\mathbf{66.35} \pm 0.12$	60.34 ± 0.22	62.41 ± 0.24	63.68 ± 0.15
CheXpert(AUC)	SN + CNN	79.15 ± 0.26	80.74 ± 0.62	79.53 ± 0.28	$\mathbf{80.99} \pm 0.08$	80.72 ± 0.35
CIFAR10(ACC)	SN + CNN	55.18 ± 0.28	$\mathbf{68.29 \pm 0.17}$	65.97 ± 0.13	66.26 ± 0.11	66.45 ± 0.45

A.2 Augmentation Multiplicity

Table A.3: Studying the impact of augmentation multiplicity, we find that it consistently improves performance for CIFAR-10. However, looking at EyePACS and CheXpert, we observe inconsistent behavior, except that it generally seems to reduce performance with EyePACS. "Best so far" refers to the baseline plus all previously effective ablations listed in Table A.4

Dataset	Model	Best So Far	+Augmult(4)	+Augmult(8)
EyePACS	SN + CNN	**66.35 ± 0.12**	65.69 ± 0.42	65.36 ± 0.19
EyePACS	B/16 + L	**69.52 ± 0.08**	69.04 ± 0.05	68.99 ± 0.11
EyePACS	B/16 + TLNN	**69.55 ± 0.06**	69.24 ± 0.08	69.35 ± 0.2
EyePACS	G/14 + L	72.04 ± 0.03	71.45 ± 0.37	71.36 ± 0.24
EyePACS	G/14 + TLNN	73.07 ± 0.05	**73.07 ± 0.05**	73.02 ± 0.07
CheXpert	SN + CNN	82.11 ± 0.29	81.41 ± 0.3	81.39 ± 0.41
CheXpert	B/16 + L	78.42 ± 0.07	78.58 ± 0.04	**78.65 ± 0.1**
CheXpert	B/16 + TLNN	78.86 ± 0.1	78.91 ± 0.04	**78.97 ± 0.12**
CheXpert	G/14 + L	81.76 ± 0.05	**82.04 ± 0.14**	81.9 ± 0.19
CheXpert	G/14 + TLNN	81.8 ± 0.06	81.06 ± 0.04	81.03 ± 0.1
CIFAR-10 (ACC)	SN + CNN	68.96 ± 0.26	69.07 ± 0.2	**69.16** ± 0.08
CIFAR-10 (ACC)	B/16 + L	99.75(94.57)	99.76(94.68)	**99.77(94.76)**
CIFAR-10 (ACC)	B/16 + TLNN	99.76(94.55)	99.78(94.75)	**99.79(94.81)**

A.3 All Ablations Together

We observe an almost consistent improvement in performance for CIFAR-10, whereas this pattern is not observed with the other datasets.

Table A.4: Studying all ablations together, we observe an almost consistent improvement in performance for CIFAR-10, whereas this pattern is not observed with the other datasets.

Dataset	Model	Best Norm	+Larger Batch	+WS	+Best Augmult	+EMA
EyePACS	SN+CNN	66.35 ± 0.12	66.24 ± 0.23	65.81 ± 0.06	65.69 ± 0.42	$\mathbf{66.65 \pm 0.16}$
EyePACS	B/16+L	$\mathbf{69.52 \pm 0.08}$	68.95 ± 0.24	-	69.04 ± 0.05	69.09 ± 0.12
EyePACS	B/16+TLNN	$\mathbf{69.55 \pm 0.06}$	69.16 ± 0.28	-	69.24 ± 0.08	69.3 ± 0.21
EyePACS	G/14+L	$\mathbf{72.04 \pm 0.03}$	71.59 ± 0.24	-	71.45 ± 0.37	71.29 ± 0.17
EyePACS	G/14+TLNN	72.66 ± 0.03	$\mathbf{73.07 \pm 0.06}$	-	73.07 ± 0.05	73 ± 0.03
CheXpert	SN+CNN	80.99 ± 0.08	81.54 ± 0.34	82.11 ± 0.29	81.41 ± 0.3	$\mathbf{82.24 \pm 0.22}$
CheXpert	B/16+L	78.35 ± 0.05	78.42 ± 0.07	-	$\mathbf{78.65 \pm 0.1}$	$\mathbf{78.65 \pm 0.04}$
CheXpert	B/16+TLNN	78.86 ± 0.1	78.86 ± 0.1	-	78.97 ± 0.12	$\mathbf{79.01 \pm 0.1}$
CheXpert	G/14+L	81.6 ± 0.01	81.76 ± 0.05	-	$\mathbf{82.04 \pm 0.14}$	82.01 ± 0.05
CheXpert	G/14+TLNN	81.8 ± 0.06	81.04 ± 0.16	-	81.06 ± 0.34	81.14 ± 0.26
CIFAR10	B/16+L	99.75(94.57)	99.74(94.49)	-	**99.77(94.76)**	99.76(94.67)
CIFAR10	B/16+TLNN	99.75(94.51)	99.76(94.55)	-	**99.79(94.81)**	99.78(94.76)
CIFAR10	SN+CNN	68.29 ± 0.17	66.47 ± 0.31	68.96 ± 0.26	$\mathbf{69.16 \pm 0.08}$	68.07 ± 0.24

A.3.1 CIFAR-10

Table A.5: Studying all ablations for CIFAR-10.

Model	Best Norm	+Larger Batch	+WS	+Best Augmult	+EMA
B/16+L	99.75(94.57)	99.74(94.49)	-	**99.77(94.76)**	99.76(94.67)
B/16+TLNN	99.75(94.51)	99.76(94.55)	-	**99.79(94.81)**	99.78(94.76)
SN+CNN	68.29 ± 0.17	66.47 ± 0.31	68.96 ± 0.26	$\mathbf{69.16 \pm 0.08}$	68.07 ± 0.24

A.3.2 CheXpert

Table A.6: Studying all ablations for CheXpert.

Model	Best Norm	+Larger Batch	+WS	+Best Augmult	+EMA
SN+CNN	80.99 ± 0.08	81.54 ± 0.34	82.11 ± 0.29	81.41 ± 0.3	$\mathbf{82.24 \pm 0.22}$
B/16+L	78.35 ± 0.05	78.42 ± 0.07	-	$\mathbf{78.65 \pm 0.1}$	$\mathbf{78.65 \pm 0.04}$
B/16+TLNN	78.86 ± 0.1	78.86 ± 0.1	-	78.97 ± 0.12	$\mathbf{79.01 \pm 0.1}$
G/14+L	81.6 ± 0.01	81.76 ± 0.05	-	$\mathbf{82.04 \pm 0.14}$	82.01 ± 0.05
G/14+TLNN	81.8 ± 0.06	81.04 ± 0.16	-	81.06 ± 0.34	81.14 ± 0.26

A.3.3 EyePACS

Table A.7: Studying all ablations for EyePACS.

Model	Best Norm	+Larger Batch	+WS	+Best Augmult	+EMA
SN+CNN	66.35 ± 0.12	66.24 ± 0.23	65.81 ± 0.06	65.69 ± 0.42	$\mathbf{66.65 \pm 0.16}$
B/16+L	$\mathbf{69.52 \pm 0.08}$	68.95 ± 0.24	-	69.04 ± 0.05	69.09 ± 0.12
B/16+TLNN	$\mathbf{69.55 \pm 0.06}$	69.16 ± 0.28	-	69.24 ± 0.08	69.3 ± 0.21
G/14+L	$\mathbf{72.04 \pm 0.03}$	71.59 ± 0.24	-	71.45 ± 0.37	71.29 ± 0.17
G/14+TLNN	72.66 ± 0.03	$\mathbf{73.07 \pm 0.06}$	-	73.07 ± 0.05	73 ± 0.03

Appendix B

Additional Details

B.1 Hardware Details

We used a collection of "NVIDIA A100 80GB PCIe"s provided by the University of Waterloo and Microsoft Azure.

B.2 Dataset License

Accessing the CheXpert dataset requires agreeing to a research use agreement on the Stanford website [39].

B.3 Code & Leaderboard

All codes for this project can be found at github.com/mshubhankar/DP-Benchmarks/. The leaderboard can also be found at private-machinelearning.github.io