The Thesis Committee for Charles Nimo
certifies that this is the approved version of the following thesis:

Deep Learning for Medical Imaging in Developing Nations

SUPERVISING COMMITTEE:

Yuke Zhu, Supervisor

Ying Ding

Deep Learning for Medical Imaging in Developing Nations

by

Charles Nimo

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computer Science

The University of Texas at Austin

May 2023

Dedication

I dedicate this thesis to my family, friends, and all those who supported me throughout my education. Thank you for being a part of this journey and helping me see it through to the end.

Epigraph

"[AI] is here to collaborate, to augment, to enhance human lives and productivity and make everybody's life better. And related to that, is to democratize A.I. in a way that everybody gets benefit. Not just a few, or a selected group."

—Dr. Fei-Fei Li

Acknowledgments

With heartfelt gratitude, I express my sincere appreciation to my esteemed supervisor **Prof. Ying Ding** for her invaluable guidance, effective supervision, and enlightenment throughout this project. I would also like to extend my special thanks to **Prof. Yuke Zhu** for graciously dedicating his valuable time to serve as the second reader for my thesis.

Furthermore, I am deeply grateful to the entire faculty of the Department of Computer Science at the University of Texas at Austin for their unwavering cooperation and support.

Abstract

Deep Learning for Medical Imaging in Developing Nations

Charles Nimo, MSCompSci
The University of Texas at Austin, 2023

SUPERVISOR: Yuke Zhu

Deep learning research and innovation have primarily been focused on high-income countries with abundant imaging data, IT infrastructures, local equipment, and clinical expertise. While the application of Deep Learning (DL) in medical imaging has gained popularity, particularly for its ability to perform on par with medical experts and bring new promises to the field o f m edicine, p rogress i n limited-resource environments where medical imaging is crucial has been relatively slow. For instance, in Sub-Saharan Africa, the rate of perinatal mortality, which refers to baby deaths during pregnancy or the first w e ek d u e t o h e althcare/maternal i s sues, i s v e ry high due to limited access to antenatal screening. In these countries, deep learning models could be implemented to help clinicians acquire fetal ultrasound planes for the diagnosis of fetal abnormalities. Although the latest deep learning models have been able to identify standard fetal planes, there is no evidence of their ability to generalize in settings with limited resources, such as areas with restricted access to high-end ultrasound equipment and ultrasound data, or different populations.

How can breakthroughs in medical deep learning research be disseminated to the global community? Moreover, how can individuals outside of America benefit from and leverage its value? In order for deep learning models to be adopted in developing countries, there is a need for greater efficiency, an d we al so re quire mo re ro bust and

privacy-preserving models to make them practical. With these questions in mind, my thesis centers on the use of deep learning in healthcare for developing nations. Specifically, I explore and propose efficient, privacy-preserving, and robust machine learning techniques to enhance the efficacy of deep learning models in healthcare. Additionally, I conduct a review of the current state of healthcare in developing regions around the world and consider how deep learning can be utilized to improve patient outcomes and support clinicians.

Table of Contents

List of Tables

Chapter 1: Introduction

1.1 Machine Learning for the Developing World

Approximately six billion people, or four-fifths of the world's population, reside in developing countries. Researchers in fields such as sociology, ecology, statistics, and economics have long studied the unique challenges faced by people in these areas. With the growing interest in machine learning research, researchers are increasingly adopting machine learning methodologies to address global development challenges. For instance, deep learning models can provide expert decision support for medical personnel in regions where resources are scarce. However, significant challenges exist when it comes to applying deep learning in these areas. Limitations in data availability, computational capacity, and internet accessibility are more prevalent in developing countries when compared to developed ones.

1.2 What opportunities can machine learning in healthcare bring to developing countries?

Machine learning drives growth through intelligent automation of the workforce, labor augmentation, and creating new opportunities for skills, business ideas, and services. In developing countries, healthcare systems often face chronic shortages of medical workers, and machine learning applications can potentially fill this gap. One such application involves machine learning-based diagnostic technology for diagnostic testing. Machine learning models can also complement highly trained and expensive expertise by analyzing medical images (Ranschaert, 2018). For instance, machine learning algorithms have shown comparable accuracy to dermatologists in classifying skin cancer, making mobile devices embedded with deep neural networks a possible tool to extend the reach of dermatologists beyond clinics (Miller and Brown, 2018) (Esteva et al., 2017). Natural language processing can also extract pneumonia-

related features from chest X-ray reports, which could help antibiotic assistant systems alert physicians to the need for anti-infective therapy. Medical practice requires managing vast amounts of data related to human physiology, imaging, and more. Assimilating and analyzing this data in a holistic and accurate manner is crucial for decision-making, and machine learning could be an invaluable complementary tool for clinicians.

1.3 Machine Learning in radiology

Machine learning techniques have been employed for several decades to detect, diagnose, classify, and assess the risk of breast cancer. The development of these techniques has resulted in remarkable disease prediction rates of up to 90.5%, indicating that machine learning models are fast, reliable, and risk-free, making them a valuable asset for physicians.

Recently, accurate deep learning models for breast cancer diagnosis have emerged as crucial tools that enhance clinical efficiency by reducing interpretation time for radiologists. Notable radiology projects using machine learning include the 'Knee OA staging' project (Thomas et al., 2020), developed at Stanford University, which aimed to automatically quantifying the severity of knee osteoarthritis from X-ray images. Various machine learning algorithms, such as SVM and region convolutional neural networks, have been explored to achieve the same goal.

Moreover, researchers have applied machine learning methods to neuroimaging data to assist with timely stroke diagnosis. For instance, Griffis et al. (Griffis et al., 2016) employed naïve Bayes classification to identify the stroke lesion in T1-weighted MRI, and the results were comparable to those of human experts manually delineating lesions. Additionally, machine learning-powered technologies incorporating multifeature analysis using diffusion-weighted imaging, magnetic field correlation, fMRI, and volumetrics have shown promise in accurately detecting and classifying traumatic brain injury in emergency settings (Lui et al., 2014). Machine learning-based health-

care solutions like these could improve productivity and healthcare outcomes for patients, and the time is right for developing countries to embrace machine learning to overcome present and future challenges in healthcare.

1.4 Creating a road-map to bridge the gap between machine learning and healthcare in developing countries

Machine learning is poised to revolutionize radiology and healthcare, offering unparalleled potential to augment and enhance the care provided to patients. By leveraging the exponential growth of available data, machine learning can help overcome the limitations of human processing power and enable more meaningful medical decision-making. Its ability to utilize individual data to diagnose diseases and create personalized treatment plans makes it a vital tool for modern healthcare.

It is high time for developing nations to take the reins of the nascent field of machine learning research and application. While the developed world has spearheaded this movement thus far, mastering machine learning is sure to bestow unparalleled dominance in all industry sectors. Despite not enjoying the same advantages in research and resources, developing countries possess a vast engineering talent pool, a burgeoning startup culture, and an abundance of untapped data. With their entrepreneurial drive and unbridled ambition, these nations are poised to help businesses extract value from real-time data and carve out a niche in an increasingly machine learning-dominated landscape (Noronha, 2018). To ensure smooth transition and implementation of machine learning in healthcare for developing countries requires strategic positioning and ethical considerations.

1.5 Opportunities for advancing healthcare through machine learning techniques

The assumptions underlying machine learning technology are tailored to developed countries, encoding the cultural and infrastructural conditions unique to these

regions. It is not surprising that the same tools may not yield optimal results when implemented in areas that do not meet these criteria. Nevertheless, machine learning has immense potential to bridge the development gap, particularly in low-resource settings lacking highly skilled experts. To fully realize this potential, we must modify our research approaches to account for the unique conditions of these developing regions. Overcoming common limitations faced in developing countries can inspire cutting-edge machine learning research.

This thesis proposes several machine learning techniques and concepts to address medical challenges in the developing world. While not exhaustive, these techniques provide a breadth of technical challenges that can inspire future novel research in various machine learning disciplines, such as efficient machine learning, machine learning robustness, learning with small or imperfect data, and privacy-preserving machine learning.

Chapter 2: Learning with Limited Memory and Computation

Medical imaging is a major anchor of clinical decision making and an integral part of healthcare. The information extracted from medical images is used in many areas such as detection, diagnosis, intervention, and treatment planning. Although medical imaging is essential to several clinical tasks, there is an increasing shortage of radiologists to interpret complex medical images, especially in developing nations. In fact, studies have shown that there exists a worldwide shortage of qualified radiologists to read, interpret, and report theses images (Rimmer, 2017) (Nakajima et al., 2008). The number of radiologists cannot keep up with the fast growth of the volume of images. Consequently, the high workload this causes leads to errors in diagnosis due to human fatigue, unacceptable delays in reporting, and stress and burnouts in radiologists. On the other hand, overdue or incorrect treatment would harm the patient. Additionally, lack of access to imaging services for diagnosis will consequently lead to many patients being denied of diagnostic treatment and preventative treatment. For this reason, there is a clear need for reliable, precise, and efficient automated methods to alleviate the growing burden on medical practitioners and improved accessibility of reliable treatment for patients' health.

Machine learning models have shown remarkable performance in automated evaluation of medical images (Hosny et al., 2018) (Suzuki, 2017) (Shen et al., 2017). For example, to diagnose various conditions from medical images, machine learning has shown to perform on par with radiologists. In recent years, there has been growing interest in the application of machine learning in medicine. In particular, deep learning is becoming an increasingly common framework for automating and standardizing important clinical tasks such as medical image analysis that would normally be subject to wide variability. However, despite growing interest, only a small number of machine learning studies have progressed to deployment in patient

care.

Research in deep learning has been focused on improving the state of the art which consequently leads to an increase in network complexity, number of parameters, memory usage, and prediction latency. For example, the GPT-3 language model consists of over 175 billion parameters (Brown et al., 2020). As large models perform well on the tasks they are trained on, unfortunately, they may not be efficient enough for direct deployment in real world clinical settings. For example, some of the challenges medical practitioners might face is the high costs in training or deploying a model. Although training could be a one-time cost, free if using a pre-trained model, deploying and letting inference run for over a long period of time could lead to expensive consumption of RAM, CPU, etc. Additionally, some deep learning enabled applications need to run real-time on IoT and smart devices to provide clinical decision support for health care workers. For a multitude of reasons including privacy, connectivity, and responsiveness, it is essential to optimize the models for the target devices.

In this chapter, we explore trade offs between model quality and model footprint to develop efficient models for direct deployment in medical imaging for thoracic diagnosis. We explore two approaches to enhance standard deep models' generalization. The first approach is to use stochastic weight averaging. While stochastic weight averaging hasn't yet been applied to sparse networks , it is well known to find the flatter minima which are widely believed to lead to stronger robustness of models. The second approach uses knowledge distillation to calibrate the well-known overconfidence of deep networks and was found to improve their standard generalization. To be perfectly clear, neither knowledge distillation or stochastic weight averaging was invented by this paper. However, by adapting them to the lottery ticket hypothesis , my aim is to complement existing studies, demonstrating that these learning techniques can be applied together with small sparse subnetworks to achieve greater performance and robustness for deep neural network chest X-ray classification .

Our experiments show that by utilizing these two techniques, we can boost the standard accuracy by 133% on chest X-Ray images.

2.1 Balancing the Tradeoffs between Model Quality and Footprint

Over the last several years there have been several attempts to decrease the number of parameters in deep learning architectures. More specifically techniques such as pruning, depth-wise separable (DS) convolutions, and filter reductions are common ways to help mitigate the overparameterization of deep neural networks for 3D medical image segmentation. (Lecun et al., 1989). (Ye et al., 2018) We focus on neural network pruning which is the kind of compression that was used to develop the lottery ticket hypothesis. We also present previous work on learning techniques used during model training.

2.1.1 Reducing Model Footprint

Compression techniques are algorithms that seek to optimize the model's architecture to achieve a more efficient representation of one or more layers in a neural network. With the efficiency goal of model optimization, compression algorithms can be used to optimize the model for one or more of the footprint metrics such as model size, inference latency, etc. Additionally in certain cases, if the model is over parameterized, these techniques can be used to improve model generalization.

Proposed by Lecun et. al, the goal of pruning is to remove the redundant connections within a neural network (Lecun et al., 1989). Given a neural network, where the input is represented by X and a set of parameters is represented with W, pruning is a technique to determine a minimal subset W' such that the rest of the parameters are pruned or in other words set to 0, while ensuring that the quality of the model is maintained above the desired threshold. Once pruning is complete, then we can say that the network has been made sparse, where sparsity is quantified as

the ratio of the number of parameters that were pruned to the number of parameters in the original network ($s = (1 - \frac{\theta'}{\theta})$). The greater the sparsity, the lesser the number of non-zero parameters in the pruned networks.

Frankle et al.'s introduced the Lottery Ticket hypothesis, their work takes a different approach to pruning and proposed that within every large network lies a smaller network that can be extracted with the original initialization of its parameters, then retrained on its own to match or exceed the performance of the larger network.

2.1.2 Improving Model Quality

Improving model quality can be achieved through the use of certain learning techniques. Learning techniques aim to to train a model differently in order to achieve better quality metrics such as accuracy, F1 score, precision, etc. while also allowing supplementing or replacing the traditional supervised learning. The balance in model quality can be traded off for a smaller footprint by reducing the number of parameters or layers in the model and maintaining the same baseline quality with a smaller model. Learning techniques are applied during training, without impacting the inference of the model.

Hinton et al, investigated how smaller networks, students, are to taught to extract the same behavior from larger models, teachers by trying to replicate it's outputs at every level. The larger teacher model is used to generate soft-labels on existing labeled data. The soft-labels are then used to assign a probability to each class. The intuition behind this approach is that the soft-labels capture the relationship between the different classes which the model can learn from. The smaller student network learns to minimize the cross-entropy loss on the soft labels, in addition to the original ground-truth hard labels. Given that the probabilities of the incorrect classes might be very small, the logits are scaled down by a 'temperature' value ≥ 1.0 so that the distribution is softened. In the study, Hinton et al.; were able to nearly match the performance of a 10 model ensemble for speech recognition

task using a single distilled model. Urban et al. introduced a study demonstrating that distillation notably improves the the accuracy of shallow student networks as small as an MLP with one hidden layer on tasks like CIFAR-10. Sanh et al. use the distillation loss for compressing a BERT model. Their model maintains 97% of the performance of BERT-Base while being 40% smaller and 60% faster on CPU.

Another idea that has been proposed recently and received a lot attention is stochastic weight averaging (SWA) (Izmailov et al., 2018). The main idea behind SWA is that it enforces thew eight smoothness, by averaging multiple checkpoints along the training trajectory. SWA has shown that it leads to much flatter solutions than SGD, is easy to implement, and improves generalization performance for deep neural networks, with almost no computational overhead. SWA has been adopted various tasks such as semi-supervised leaning (Athiwaratkun et al., 2018), Bayesian inference (Maddox et al., 2019), and low-precision training (Yang et al., 2019). In this paper, we introduce SWA to sparse models for the first time, in order to smooth the weights and find flatter minima that may lead to improvements of robust generalization.

Hinton et al (Hinton et al., 2015), investigated how smaller networks, students, are to taught to extract the same behavior from larger models, teachers by trying to replicate it's outputs at every level. The larger teacher model is used to generate soft-labels on existing labeled data and then the soft-labels are then used to assign a probability to each class. The main idea behind this approach is that the soft-labels capture the relationship between the different classes from which the model can learn from. The smaller student network learns to minimize the cross-entropy loss on the soft labels, in addition to the original ground-truth hard labels. Considering that the probabilities of the incorrect classes might be very small, the logits are scaled down by a 'temperature' value ≥ 1.0 so that the distribution is softened. In the study, Hinton et al.; were able to nearly match the performance of a 10 model ensemble for speech recognition task using a single distilled model. Urban et al. introduced a study demonstrating that distillation notably improves the the accuracy of shallow student networks as small as an MLP with one hidden layer on tasks like CIFAR-10.

Sanh et al. use the distillation loss for compressing a BERT model. Their model maintains 97% of the performance of BERT-Base while being 40% smaller and 60% faster on CPU.

Another idea that has been proposed recently and received a lot attention is stochastic weight averaging (SWA) (Izmailov et al., 2018). The main idea behind SWA is that it enforces thew eight smoothness, by averaging multiple checkpoints along the training trajectory. SWA has shown that it leads to much flatter solutions than SGD, is easy to implement, and improves generalization performance for deep neural networks, with almost no computational overhead. SWA has been adopted various tasks such as semi-supervised leaning (Athiwaratkun et al., 2018), Bayesian inference (Maddox et al., 2019), and low-precision training (Yang et al., 2019). In this paper, we introduce SWA to sparse models for the first time, in order to smooth the weights and find flatter minima that may lead to improvements of robust generalization.

Chapter 3: The Privacy-Utility Tradeoff in medical machine learning

3.1 Differential Privacy

Machine learning models rely heavily on the quantity and quality of data available during training. This poses a significant challenge in domains like medical imaging, where high-quality data is sparse and the use of data is restricted due to ethical and regulatory considerations. Medical data is often sensitive, and sharing it is limited by ethical requirements. As a result, the progress of algorithms that generalize well is impeded, and their widespread deployment in clinical settings is prevented. However, given that state-of-the-art computer vision models are typically trained on large datasets like ImageNet (Deng et al., 2009), it's clear that access to more data will be necessary for most deep learning applications in medical imaging to succeed. Privacy-preserving machine learning is a rapidly growing research area that aims to balance data utilization and protection using privacy-enhancing techniques. One of the most promising techniques is federated learning, which enables a confederation of clients to train machine learning models in a decentralized manner without sharing the raw data (Sheller et al., 2020). However, research has shown that federated learning alone may not be enough to preserve privacy. In medical imaging, it could lead to significant privacy loss for patients. Moreover, some prior works have shown that federated learning without additional privacy-enhancing techniques could be reverse-engineered to recover high-fidelity images that encode sensitive diagnostic information about patients. Therefore, comprehensive solutions are needed that can provide quantifiable guarantees to all parties involved.

Differential Privacy (DP) has emerged as the gold standard for ensuring privacy (Dwork and Roth, 2014). DP is a powerful technique that was initially developed in statistics and has been increasingly utilized in machine learning research to safeguard datasets from being exposed based on their outcomes. By adding noise to

inputs, outputs, ground truth labels, or even models themselves, DP ensures that adversaries cannot predict or infer individual records after deploying machine learning models in public or presenting only the results. In their research, Abadi et al. (Abadi et al., 2016) successfully applied DP to training deep neural networks using differentially private stochastic gradient descent (DP-SGD). However, subsequent works, including the authors themselves, have observed that DP-SGD negatively impacts the utility of resulting models, leading to a known side effect called the privacy-utility trade-off (Avent et al., 2020). To enable real-world deployment of privacy-preserving deep learning in medical imaging and other areas, it is crucial to address this trade-off by introducing training methodologies that maintain privacy guarantees without compromising the resulting model's quality.

3.2 DiWA

In computer vision, the conventional method to maximize model accuracy is to train deep neural networks with various hyperparameters and pick the individual model which performs best on a held-out set. Typically, the best performing models are those that are pre-trained on a large dataset before being fine-tuned on data from the target task(Donahue et al., 2013; Yosinski et al., 2014; Razavian et al., 2014; Girshick et al., 2014). Most recently, Rame et al. proposed Diverse Weight Averaging (DiWA), a method aimed at increasing the functional diversity across averaged models (Ramé et al., 2023). DiWA takes the averages of weights obtained from several independent training runs of multiple models. The authors show that models obtained from different runs are more diverse than those collected along a single run and that this diversity leads to better generalization. In addition, DiWA introduces no additional inference cost, which is a known key limitation in comparison to ensemble techniques. Inspired by this work, I propose using a DiWA-inspired training regime with DP-SGD to achieve better utility for differentially private deep neural networks. I successfully demonstrate that this new training regime leads to

better generalization for private models.

Chapter 4: Experiments and Results

4.1 Deep Learning for healthcare - Efficient Machine Learning

In this chapter, I will elucidate the experimental settings and techniques employed to identify subnetworks, followed by a comprehensive discussion and analysis of the results.

4.1.0.1 Network

I used the official DenseNet-121 architecture as the original (unpruned) dense networks. The DenseNet model was selected because prior studies Rajpurkar et al. (2017) have found that it works well to predict Pneumonia disease. $f(x;\theta)$ defines the output of the DenseNet model with parameters $\theta \in \mathbb{R}^d$ and for input images x. Correspondingly, sparse subnetworks extracted from the dense model θ are represented as $m \odot \theta$ where $m \in \{0,1\}^d$ is a pruning binary mask and $\odot$ represents the element-wise product. Sparsity is quantified as the number of parameters that were removed due to pruning, to the total number of parameters in the original network ($s = (1 - \frac{\theta'}{\theta})$).

4.1.0.2 Pruning Methods

In order to find the sparse subnetworks $f(x; m \odot \theta)$, I leveraged *the lottery ticket hypothesis* (LTH) with *one-shot magnitude pruning* (OMP), as it is more efficient, by removing a portion of the weights with the globally smallest magnitudes. Following he LTH routine, we prune the network to 9% of the remaining weight and rewind model weights to the same random initialization. In order to evaluate whether a subnetwork is matching on the original task, we train it using $\mathcal{A}_t^{\mathcal{T}}$, where the network is trained to completion on a task $\mathcal{T}$ for t steps. A subnetwork is said to be matching

for an algorithm $\mathcal{A}_t^{\mathcal{T}}$ if training $f(x; m \odot \theta)$ with algorithm $\mathcal{A}_t^{\mathcal{T}}$ results in evaluation metric that asseses the task-specific performance on task $\mathcal{T}$ no lower than training $f(x; \theta_0)$ with algorithm $\mathcal{A}_t^{\mathcal{T}}$. That is:

$$\epsilon^{\mathcal{T}}(\mathcal{A}_t^{\mathcal{T}}(f(x; m \odot \theta))) \geq \epsilon^{\mathcal{T}}(\mathcal{A}_t^{\mathcal{T}}(f(x; \theta_0))))$$

4.1.0.3 Datasets and Pre-Processing

In order to evaluate the performance of both the original dense network and sparse subnetwork we conducted experiments on the public kaggle dataset, RSNA Pneumonia Detection Challenge. Originally, the dataset contains 30,227 frontal-view images in Dicom format and three classes whee 31.61% of the images are of the lung opacity (pneumonia), 39.11% no lung opacity, and 29.28% normal images. However, the no-lung opacity class was removed to create a new two class dataset. This resulting dataset contained 14,863 images where several duplicates that were found were dropped. We used 60% for training and fine-tuning, 20% for validation and 20% for testing. For both the original network and matching subnetwork, we iterated the training and fine-tuning process for 100 epochs with batch size 16 and early stopped if the loss did not decrease. We compare the AUROC (area under the receiver operating characteristic curve) of the subnetwork immediately after pruning its parameters to the AUROC of the subnetwork after training and fine-tuning it.

4.1.0.4 Experiments and Analysis

In this section, I comprehensively investigate the performance of sparse neural networks from LTH with the following learning techniques *(i)* stochastic weight averaging, *(ii)* knowledge distillation with weight rewinding. *(iii)*

4.1.0.5 Learning with SWA

Stochastic Weight Averaging (SWA) was introduced in training of the sparse subnetwork to lead to better robust model performance. SWA averages the weights along the same optimization trajectory with one signle run. It is known to find much flatter solutions than SGD and is also very straightforward to implement, enhances standard generalization, and introduces almost no computational overhead. Following Weng et al. (2019) Izmailov et al., we apply SWA such that:

$$W_{SWA} = \frac{W_{SWA} \cdot n + W}{n + 1}$$

where W_{SWA} is the computed average network weight, W is the current network weight, and n is the number of past checkpoints to be averaged.

4.1.0.6 Learning with Knowledge Distillation and Weight Rewinding

I studied the effectiveness of knowledge distillation to train the sparse subnetwork on the outputs of the original dense network, following a student-teacher paradigm. Hinton et al., Hinton et al. (2015) made the observation that knowledge distillation is a very successful technique for knowledge transfer between classifiers. Inspired by the success of rewinding, Ma et al., Ma et al. (2021) propose knowledge distillation tickets (KD Ticket), where they observe that knowledge distillation tickets applied together with weight rewinding yields state-of-the-art performance in large-scale ticket findings. Their insight comes from the observation that the weights at early training points hold more detailed information of the original training process to maintain the accuracy. Following the routine as described in Ma et. alMa et al. (2021)'s work , we adopt the same technique of selecting the best rewind point between 18% and 29% for the training phase of the original dense network. More specifically, we rewind the remaining weights of the pruned network to the specific

early training point, and retrain the pruned subnetwork suing the soft labels of the original network.

Label	AUC
Pneumonia	0.9898
Normal	0.9897

Table 4.1: AUC score of binary classifier before prunning

Label	AUC
Pneumonia	0.4247
Normal	0.4004

Table 4.2: AUC score of binary classifier after prunning

Label	AUC
Pneumonia	0.9897
Normal	0.9898

Table 4.3: Binary Classifier retrained with SWA

Label	AUC
Pneumonia	0.9853
Normal	0.9851

Table 4.4: Binary Classifier Retrained with KD + Rewind

4.1.0.7 Analysis

From the results shown in tables 4.1, 4.2, 4.3, 4.4. We see that fine-tuning the sparse networks leads AUC scores comparable to the original dense network. SWA is shown to outperform KD + rewind fine-tuning. There are several reasons why SWA yields better generalization than KD + rewind technique. One of the reasons being is that SWA averages the weights of the model at different epochs, which can help reduce the impact of noisy or misleading updates during training. Also, SWA can

also reduce overfitting by preventing the model from getting stuck in local minima. Since it averages the weights at different epochs, SWA can more efficiently smooth out the fluctuations in the loss landscape and help the model escape from suboptimal local minima.

4.2 Deep Learning for healthcare - Privacy Preserving Machine Learning

In this chapter, I will elucidate the experimental settings and techniques employed to develop a new training regime to alleviate the trade offs between privacy and accuracy using multiple fine-tunings of the same model on hybrid tasks.

4.2.0.1 Network

I use the official PyTorch implementation of the pre-trained ResNet-20 architecture. The ResNet-20 model was selected because prior studies Rajpurkar et al. (2017) have used this model when benchmarking the performances of differenitally private models.

4.2.0.2 Datasets and Pre-Processing

I conduct experiments on the MNIST and CIFAR-10 datasets.

4.2.0.3 Evaluated Algorithms

I use the current state of the art differential privacy algorithms to serve as baselines against my approach: PATE - private aggregation of teacher ensembles (Papernot et al., 2018), DPSGD - differentially private stochastic gradient descent Abadi et al. (2016), RGP - Reparameterized Gradient Perturbation Yu et al. (2021).

Algorithm	epsilon = 8
PATE	38.44
DPSGD	61.81
RGP	67.81
SR	68.78

Table 4.5: Test Accuracy of 'SR' in comparision to other DP algorithms on CIFAR-10 Dataset

Algorithm	epsilon = 8
PATE	98.51
DPSGD	94.72
RGP	96.63
SR	98.83

Table 4.6: Test Accuracy of 'SR' in comparision to other DP algorithms on MNIST Dataset

4.2.0.4 Experiments and Analysis

In this section, I comprehensively investigate the performance of my approach referred to as "SR" against the following baselines, PATE - private aggregation of teacher ensembles (Papernot et al., 2018), DPSGD - differentially private stochastic gradient descent Abadi et al. (2016), RGP - Reparameterized Gradient Perturbation Yu et al. (2021). For this approach, I first downloaded a pretrained ResNet-20 model. Secondly, I fine-tuned the pretrained model for twenty different and independent training runs on both the CIFAR-10 and MNIST dataset each, using linear probing, which learns the classifier to serve as a shared initialization for all training runs, this was used instead of random initialization, which may distort the features. Linear probing fine-tunes the encoder and the classifier together in the 20 subsequent runs. The last step of the approach is to choose which weights to average. To select which model weights to average, I used a greedy approach which adds the top weights of models to average, ranked in decreasing order of validation accuracy. As shown in the following tables 4.5, 4.6, my approach referred to as 'SR' achieves improved test

accuracy of private models in comparison to the current state of the art methods for differentially private deep neural networks.

Chapter 5: Conclusion

In this thesis, I perform a comprehensive evaluation for the effectiveness of the use of learning techniques on sparse subnetworks to achieve better test accuracy. I beleive that these compelling results yield several in-depth insights of understanding network pruning and lottery ticket hypothesis. Further more, I also believe that are findings endorse the wider adoption of sparse subnetworks in replacement of dense networks. Our future works will extend this study to include other kinds of model compression methods such as along with learning techniques such as unsupervised learning and adversarial machine learning to further improve robustness.

Additionally, I demonstrate the success of weight averaging based on similarity with ensembling can improve utility for differential privacy training of deep neural networks. Through the use of diverse pretrained models, the averaged weights of independent training runs is efficient when models can be connected linearly in the weight space during differential private training.

I believe this scientific research poses great societal impact in the healthcare domain, especially for developing countries. With the assistance of this study, it may be more possible to establish efficient, robust, and private sparse deep neural networks with reduced energy and financial costs to provide clinical decision support in medical imaging diagnosis for medical practitioners in areas where access to medical specialists is scarce.

Works Cited

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 SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318. URL `https://doi.org/10.1145%2F2976749.2978318`.

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average, 2018. URL `https://arxiv.org/abs/1806.05594`.

Brendan Avent, Javier Gonzalez, Tom Diethe, Andrei Paleyes, and Borja Balle. Automatic discovery of privacy-utility pareto fronts, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for

generic visual recognition, 2013.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, Feb 2017. ISSN 1476-4687. doi: 10.1038/nature21056. URL https://doi.org/10.1038/nature21056.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. doi: 10.1109/CVPR.2014.81.

Joseph C. Griffis, Jane B. Allendorfer, and Jerzy P. Szaflarski. Voxel-based gaussian naïve bayes classification of ischemic stroke lesions in individual t1-weighted mri scans. *Journal of Neuroscience Methods*, 257:97–108, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence Schwartz, and Hugo Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18, 05 2018. doi: 10.1038/s41568-018-0016-5.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2018. URL https://arxiv.org/abs/1803.05407.

Yann Lecun, John Denker, and Sara Solla. Optimal brain damage. volume 2, pages 598–605, 01 1989.

Yvonne Lui, Yuanyi Xue, Damon Kenul, Yulin Ge, Robert Grossman, and Yao Wang. Classification algorithms using multiple mri features in mild traumatic brain injury. *Neurology*, 83, 08 2014. doi: 10.1212/WNL.0000000000000834.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Good students play big lottery better, 01 2021.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning, 2019. URL https://arxiv.org/abs/1902.02476.

D. Douglas Miller and Eric W. Brown. Artificial intelligence in medical practice: The question to the answer? *The American Journal of Medicine*, 131 (2):129–133, 2018. ISSN 0002-9343. doi: https://doi.org/10.1016/j.amjmed.2017.10.035. URL https://www.sciencedirect.com/science/article/pii/S0002934317311178.

Yasuo Nakajima, Kei Yamada, Keiko Imamura, and Kazuko Kobayashi. Radiologist supply and workload: international comparison–working group of japanese college of radiology. *Radiation medicine*, 26:455–65, 11 2008. doi: 10.1007/s11604-008-0259-2.

Vanita Noronha. Making a case for cancer research in india. *Cancer Research, Statistics, and Treatment*, 1(1), 2018. ISSN 2590-3233. URL https://journals.lww.com/crst/Fulltext/2018/01010/Making_a_case_for_cancer_research_in_India.18.aspx.

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

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization, 2023.

Erik Ranschaert. Artificial intelligence in radiology: Hype or hope? *Journal of the Belgian Society of Radiology*, 102, 11 2018. doi: 10.5334/jbsr.1632.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition, 2014.

Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ*, 359, 2017. ISSN 0959-8138. doi: 10.1136/bmj.j4683. URL https://www.bmj.com/content/359/bmj.j4683.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10 (1):12598, Jul 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-69250-1. URL https://doi.org/10.1038/s41598-020-69250-1.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. doi: 10.1146/annurev-bioeng-071516-044442. URL https://doi.org/10.1146/annurev-bioeng-071516-044442. PMID: 28301734.

Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological Physics and Technology*, 10, 07 2017. doi: 10.1007/s12194-017-0406-5.

Kevin Thomas, Łukasz Kidziński, Eni Halilaj, Scott Fleming, Guhan Venkataraman, Edwin Oei, Garry Gold, and Scott Delp. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiology: Artificial Intelligence*, 2:e190065, 03 2020. doi: 10.1148/ryai.2020190065.

Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019. doi: 10.1109/ACCESS.2019.2908991.

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Christopher De Sa. Swalp : Stochastic weight averaging in low-precision training, 2019. URL `https://arxiv.org/abs/1904.11943`.

Rongtian Ye, Fangyu Liu, and Liqiang Zhang. 3d depthwise convolution: Reducing model parameters in 3d vision tasks, 2018. URL `https://arxiv.org/abs/1808.01556`.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization, 2021.

Vita

Charles Nimo was born in Alexandria, Virginia. He graduated from Virginia Commonwealth University in 2017 with a bachelor's of science degree in Electrical and Computer engineering. He then on went on to work for Dell and Intel as a software engineer for four years.

Address: nimo@utexas.edu

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.