METHODS FOR DOMAIN-SPECIFIC FINE-TUNING FOR GENERATIVE MODELS

by
Yijing Zhang

A thesis submitted in partial fulfillment of
the requirements for the degree of

Bachelor of Science, Honors in the Major
(Computer Science)

at the
UNIVERSITY OF WISCONSIN-MADISON
2024

Superviser:
Prof. Frederic Sala

© Copyright by
Yijing Zhang
2024

Abstract

In the realm of machine learning, the need for high-quality datasets for model training is increasing, highlighting the issue of data scarcity. Generative models offer a potential solution by creating synthetic datasets. However, current generative models often struggle with out-of-domain data generation. Fine-tuning techniques can adapt generative models to new domains, but their performance when training other models remains understudied. This thesis aims to address this gap by examining how domain-specific synthetic datasets can be used for model training.

To investigate the problem, we propose a two-way evaluation pipeline, which involves training and testing two domain-specific classifier models (one trained on a synthetic dataset and tested on a real-world dataset, while the other is trained on a real-world dataset and tested on a synthetic dataset) and then compare their performance. Our experiments reveal a *trainability gap* (T-Gap) between the effectiveness of training with synthetic data and the fidelity of that data to the target domain. We introduce the fine-tune efficiency ratio (FTER) metric to quantify this gap, with values closer to one indicating a smaller gap and better model performance.

We apply this pipeline and metric to various experimental settings, focusing on techniques for data selection and architecture improvements. Results show that integrating an active-data selection architecture into the domain-adaptation model, along with cluster-based training data pre-processing, yields the best performance under the 2-way pipeline and FTER metric.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Frederic Sala, for his support, guidance, and encouragement throughout the entire journey of this thesis. His insights, constructive feedback, critical review, and patience have been instrumental in shaping this research.

Thanks to the L & S Honors Program for giving me this incredible opportunity to complete this thesis. I appreciate the opportunity to present my work at the Senior Honors Thesis Symposium 2024.

Thanks to the University of Wisconsin-Madison for supporting access to all research resources. Thanks to the computer science department. I am indebted to every staff member of the Computer Sciences Department for creating this incredible undergraduate experience here. Their expertise and commitment to academic excellence have enriched the quality of this study. Special thanks to the academic advisors for giving me guidance for this thesis.

I am grateful to my friends for their encouragement, motivation, and understanding during moments of doubt and difficulty. Their unwavering belief in my abilities has been a constant source of inspiration.

I would also like to extend my heartfelt appreciation to my family for their love, encouragement, and unwavering support throughout this journey. Their sacrifices and belief in my aspirations have been the driving force behind my achievements.

This thesis would not have been possible without the contributions of each individual mentioned above. Thank you for believing in me and being a part of this experience.

Table of Content

Chapter 1

Introduction

This thesis is concerned around the problem of data paucity and data-efficiency problem with current generative models and proposing a set of new metrics of measuring generating tasks. The thesis begins with introducing a bried background around the research problem the thesis targeted on. In this chapter, the thesis will also present an outline of the purpose of our research, including our research question, and an overview of the results of our research. We will start our thesis by discussing the background related to the topic of this thesis.

1.1 Background

Artificial intelligence has shown remarkable achievements in recent years. Generative models, in particular, poss their outstanding ability to replicate high-quality synthetic images, text and even massive data, representing a capstone of generation of machine learning models. Unlike discriminative models that focus on classification or regression tasks, generative models are machine learning models designed to understand, learn, model and generate knowledge of complex datasets. They are able learn the underlying distribution of a dataset by capturing its patterns and structures, and able to reproduce new sample from the learned distribution.

There are several approaches to structure an generative model. One common approach is using generative adversarial networks (GANs) [20]. GANs operate the training on two neural networks, generator and discriminator, which generator generates samples from random noise and discriminator learns to distinguish between the generated results and the real data inputs.

Another approach in to use the probabilistic machine learning, where the model are design in particular to learn and predict the distribution of the data and resample generated results from the learnt distributions. Several common technique to achieve this include autoregressive models, which the model predicts the probability distribution of each data point conditioned on previous points, and variational autoencoders (VAEs) [34], which learn a latent space representation of the data and then generate new samples by sampling from this latent space.

Recently, in particular, latent diffusion models [58], based on probabilistic machine learning, have offered impressive abilities to generate high-quality image gain attention. Diffusion models are learning probability distributions through an forward diffusion process of gradually destroying the data disstribution and a reverse diffusion process of restoring the data distribution [63]. Latent diffusion, in particular, learn through the forward process of adding random noise to image dataset and the reverse diffusion process of learning to denoise to produce high-quality images [58].

Such abilities in high-quality generation task has gained attention across various fields. Currently, generative models already have a wide range of applications, including text, image and speech generation, and continue to be an active area of research in machine learning and artificial intelligence. Furthermore, generative models have shown remarkable capabilities in tasks from general image generation to domain-specific medicine discovery tasks [37].

However, current powerful machine learning models trades high performance with the massive size of its training data, posing the problem of data paucity. We will discuss the problem in detail in next section.

1.1.1 Data Paucity

Generative models heavily rely on the amount of high-quality dataset to effectively learn and reproduce the data distribution they are tasked with generating. Targeting improved performance and outcome, increasingly massive scale of data are being used for training large foundation models [54]. For example, the latent diffusion models [58], one of the most powerful image generation models currently, considered among the most powerful image generation models today, draw its power by taking advantage of being trained on large-scale, multi-modal dataset LAION-5B [60], which contains about 170M high-quality image examples. Such instance underscores the "scaling law" for language models, where the more data typically yield better training results [33]. There is a growing recognition of the critical need for large-scale, high quality datasets for training machine learning models.

Nevertheless, acquiring such large-scale datasets can present significant challenges and limitations. The process of collecting, annotating and meticulous labeling demands significant financial resources and human effort. Additionally, with growing concerns surrounding data privacy, especially in the possible leaking of privacy information in generated results

[32], such problem can be more severe and present more limits in the performance of machine learning models in some specific domains. Particularly, in medical domain, where data annotation can requires massive manual effort of well-trained expertise and patient information can be rigorously protected, the dearth of high-quality, large-scale trainable data is particularly acute. Even, the data for some diseases itself can be rare in the real-world, while learning characteristics of those diseases using language models might help docter learn more about those diseases. Therefore, the deficit of high-quality datasets emerges as a a formidable obstacle [6].

To combat the problem of data paucity, generative models posses a "cheap" and zero-privacy-concern solution of synthetic data generation as alternative extensive training data for training machine learning models [65]. By generating high-quality synthetic datasets, it becomes possible to train machine learning models effectively in multiple data-deficient domains, enhancing human capabilities across various tasks.

Yet, the capabilities of current generative models, e.g. latent diffusion models [58], shows domain limitation on generating tasks constrained by the domain of their training dataset. The challenge of out-of-domain generation persist [67]. Gaps in distributions between training data and targeted dataset means the task requires transfer training between distributions [52]. While the task of out-of-domain generation typically relates to domain with significantly insufficient data, such as medical health, emerging the problem of transfer learning of out-of-domain generation tasks. Techniques addressing such distribution gaps is necessary, and one possible and popular solution is model fine-tuning [65].

1.1.2 Model Fine-tuning

Model fine-tuning is an essential technique in leveraging pre-trained models to adapt specific tasks or domains, which significantly save the costs and limits of training from scratch. For large foundation language models, fine-tuning technique usually involves taking one pre-trained large language model, typical one robust model trained on a large dataset for a general task (e.g. latent diffusion [58]), and perform some further training on a smaller task-specific/domain-specific dataset to fit the model into certain task/domain, which the method is also called the Universal Language Model Fine-tuning (ULMFiT) [27]. The process of fine-tuning often includes replacing the output layer or adding task/domain-specific training or attention layer to the pre-existing architecture, and updates the model parameter through gradient descent optimization techniques to minimize a pre-defined loss function, which usually designed to efficiently learn task/domain specific knowledge (e.g. Class-specific prior preservation loss proposed in Dreambooth model [59]). Such process enables the model to learn new distributions for specific tasks/domains and acquire task-specific knowledge, while still benefiting from preserving knowledge encoded in the pre-trained parameters. Fine-tuning process promotes more robust and efficient deep learning models [30].

Fine-tuning has found widespread application to transfer model training across various domains [24, 27]. For instance, for the task of medical image generation (Chest X-rays in particular), fine-tuned generation models (latent diffusion [58]) has proven to be effective in generating high-quality realistic images [11]. The versatility and effectiveness of fine-tuning make it a fundamental technique in modern machine learning workflows, empowering practitioners to leverage pre-existing knowledge and adapt it to novel challenges efficiently. The prospect of generating synthetic datasets as an alternative solution to address the training data paucity problem holds promise for further advancements.

However, fine-tuning techniques still remain certain restrictions and challenges. One critical aspect of fine-tuning is the balance between how much to pre-trained to preserve and how much to adapt the model to the new task. One potential problem is the distortion of pre-trained features, which can be beneficial knowledge to preserve, and underperform the specific prediction task/domain of interest [38]. Also, overfitting to the target dataset, where the model learns noise or reproducing results too similar to tuning set [26], is another common challenge. In task in image generation in particular, though current fine-tuned generative model are capable in generating high quality of synthetic images [24, 11], whether the synthetic images are capable for training purposes still remains unclear, requiring further research on the retrainability of synthetic datasets and thereby optimizing the fine-tuning process for a wide array of applications and domains.

1.1.3 Efficiency

As the realm of machine learning develops, there exists an increasing pursuit for model complexity and accuracy. Though current models has cracked a variety of complicated tasks, including text generation, image generation and speech generation, model training nowadays require incredibly large and even increasing computational and data resources. Consequently, one critical concern has emerged: efficiency.

The efficiency problem revolves around the trade-off between model performance and data and computational resource requirements. The pursuit of higher accuracy often acquire complex architecture and large model size, which also requires massive training data to build up the training for model parameters, dramatically driving up resource requirements. Though advancing hardware development continuously improving computational restrictions, computational resources are still expensive and hard to obtain. The accessibility of computational or data resources determine the practical viability of deploying machine learning models across various domains. Moreover, due to data paucity and accessibility limitations, data resources available for model training present significant challenges, and thus the task of efficiently maximizing training data utility has increasing importance [40, 1]. While better performance is desirable, in order to promote multiple usage of machine models in diverse domains, efficiency is important problem to deal with. The task of model fine-tuning is facing the same challenge. Even model fine-tuning has shown impressive results in adapting domain-specific generalization tasks [24, 11], it still remains a problem on how to fine-tune efficiently.

There are several approaches in addressing the efficiency problem in machine learning models. One avenue is compressing model architecture to reduce the size and computational complexity of neural networks while not compromising performance significantly. Another important approach in to make the model training more data-efficient. Data-efficient models are capable to learn data feature with high ability and able to perform good results with less training steps, and thus could significantly save computational and data resources [1]. There are two major approach for better data-efficiency models: one is to develop algorithmically more data-efficient architecture; the other is to get better quality data, which can be efficiently learned, into training.

1.2 Research Goals

As stated previously, in the current landscape of machine learning, one fundamental challenge is the scarcity of massive, high-quality, trainable data across multiple domains, which present a significant obstacle for the widespread application of machine learning models, particularly in specialized fields such as medical health. Despite, in theory, model fine-tuning, which is the technique of adapting robust pre-trained models to specific downstream tasks/domains by leveraging existing knowledge and adjusting model parameters to learn the features of the tuning dataset, posses a promising future solution to address the data paucity problem by generating synthetic datasets and "feed" them to train of large machine learning models, the practical implementation of model fine-tuning maintains significant challenges and concerns. The quality and trainability of the synthetic datasets still remain uncertain. The data scarcity problem continues to hinder progress, as even fine-tuning processes may require the access of sufficient amounts of diverse data for effective out-of-domain model tuning. Furthermore, the massive computational resource consumption of tuning large models with massive amount of parameters remains a critical concern. It then draws to the concern of the model efficiency problem, particularly an increasing focus on data learning efficiency of machine learning models. Therefore, robust methodologies for evaluating the quality, suitability, trainability and efficiency of synthetic datasets for generative models are essential.

Research Questions: The two major questions the thesis is focused on are:

1. How to effectively evaluate trainability performance of synthetic datasets;
2. How to/what are methods to efficiently fine-tune image generation models for out-of-domain tasks given fair trainable quality of synthetic datasets.

This thesis study aims to investigate the trainability performance of synthetic datasets and efficiency performance of different methods of domain-specific fine-tuning for generative models, especially focusing on the domain of medical imaging.

This thesis proposes a two-way evaluation pipeline of judging the trainability performance of synthetic datasets for generative models. Figure 1.1 shows the pipeline of our evaluation task. One route trains a general classifier with synthetic data and test on real-world data. Conversely, the other route trains the classifier with real-world data and test on synthetic data. By then comparing the outcomes of these two routes, we seeks to gain insights into the trainability of the synthetic datasets and the efficiency performance of the fine-tuning model.

Along with the two-way pipeline, this thesis also proposed a new metric, Fine-tune Efficiency Radio (FTER), which measures the the gap, namely Trainablity Gap (T-Gap), between the performance of synthetic datasets generated by specific models serve as training set and as testing set. We explained this gap as the lack of ability in reproducing the variance of the full underlying data distribution of the targeted domain. For future researches, we foresee more use of the FTER metric to evaluate such gap and explain such gap.

This thesis finds that an active data-selection based fine-tuning approach combined with some clustering data selection selection is a promising way to get more trainable synthetic datasets and also reaching generally data-efficient training process.

1.3 Thesis Outline

Below, we present a brief outline containing the contributions of this thesis. We will first briefly discuss related works around this topic. Then, we will outline the our methods for our research in Chapter 3. Chapter 4 will be a detailed representation of the experiments

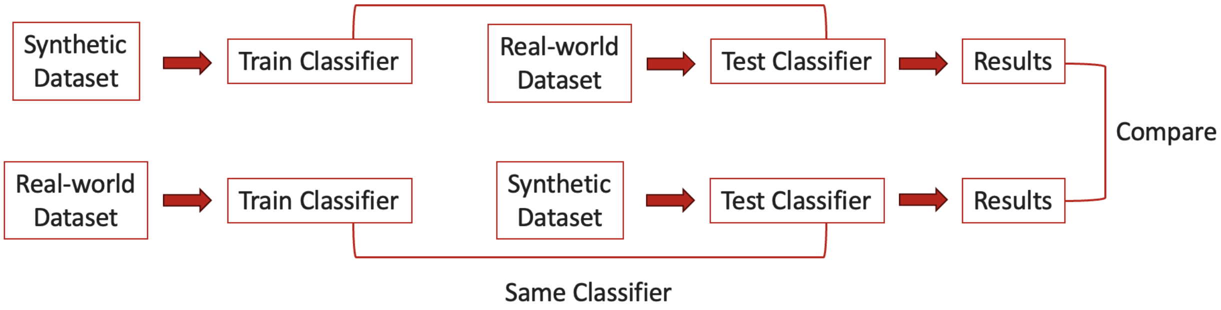


Figure 1.1: Our proposed two-way evaluation pipeline

that we did, followed by a detailed discussion of our work and contributions in Chapter 5. Our contributions and future directions for our work are summarized in Chapter 6.

Chapter 2

Literature Review

This chapter will be a literature review of related topics, including generative models, out-of-domain generation tasks, model fine-tuning and techniques to address data efficiency problem, encompassing their theoretical foundations, methodological developments, and practical applications.

2.1 Generative Model

Typically, in machine learning, models are divided into two categories: generative and discriminative. Generative models generates new data instances, while discriminative models discriminate/differentiate between different data instances. Our thesis will focus on generative models.

As briefly introduced in Chapter 1, generative models have accomplished remarkable advancements in machine learning with its the excellent capability to generate synthetic data. The foundational idea of generative model is learn and reproduce distributions of training data. More mathematically, the models aim to learn and capture the joint probability $\mathbb{P}(X, Y)$ given a set of data X and a set of labels Y (if any) [46].

There are several approaches for modeling such predictions. Classical approaches, including Gaussian mixture models (GMMs) and Markov models, attempted to represent the conditional probability of targeted data as a mixture of simpler distributions/probability. However, those classical approaches present failures in modeling high-dimensional, complex data. With the advance in the field of deep learning, modern approaches of constructing generative models include variational autoencoders (VAE) [34], generative adversarial networks (GAN) [20], flow-based generative models [35], generated pre-trained Transformers (GPTs) [53] and latent diffusion models [58]. We will discuss those architectures in more detail in the following subsections.

Generative models have found widespread applications across various domains. In computer vision, generative image models have been used for high-resolution image synthesis tasks. The impressing ability in text generative of generative models also being employed in natural language processing. In drug discovery, generative models have been utilized for molecular design and property prediction.

Overall, we foresee the potential of generative models to serve as essential tools to help enhance human ability across multiple tasks.

2.1.1 Classical Approaches

Classical approaches of modeling a generative model usually start by representing targeted data distribution/probabilities as an outcome of some distributions/probability. They are usually modeled under specific assumptions on the structure of the data, and thus present limitations based on their assumptions. Some essential classical approaches includes:

Mixture Models (MMs) are among one of the earlier approaches of generative models. It is a latent variable model that assumes some distribution over the variables conditioned on its latent space and try to represent the probability of targeted data as a mixture of some distributions over the variables. Commonly, Gaussian distributions over the variables are assumed, forming a type of mixture models called Gaussian Mixture models (GMMs).

For Gaussian Mixture models, symbolically, it predicted the probability of targeted data X: $\mathbb{P}(\mathbf{X}) = \sum_{i=1}^{K} \pi_i N(\mathbf{X}|\mu_i, \mathbf{\Sigma}_i)$, given μ_i and Σ_i as the mean vector and covariance vector of the Gaussian distributions. Data is generated by sampling data point from the estimated Gaussian distributions with differently sampled mixing coefficient pi given.

The parameters of the distributions in mixture models are usually estimated using Expectation Maximization (EM) algorithm, which involves a E-step and a M-step in its prediction. In the E-step, the algorithm represent a likelihood estimation using the current parameters sets. In the M-step, the algorithm updates parameters to that maximized the likelihood estimation.

One important feature of MMs is that they assume all datas are drawn from some distributions independently. This assumptions make MMs to perform moderately well in modeling sequential data, such as time series and speech, where the assumptions are more likely to be fulfilled. However, such assumption still present significant limitations, especially in modeling high-dimensional data [4].

Markov models are generative models based on Markovian process. They assumes a Markovian process underlying the data generation, where sequences/states depends solely on fixed number of previous states. The models aim to estimate the probability of targeted data event as an outcome of previous given sequential data stages, modeling using Markov Chain technique.

Markov models shows to be powerful tools in modeling and generating sequences of data, such as text generative where it can learn the probabilities of words or characters given some previous text sequences. By capturing the statistical dependencies between sequences and states, Markov models can effectively reproduce features of the training data.

However, Markov models present restrictions based on the Markovian assumption. Markovian process assumes data sequences are dependent to each other, which does not always hold true for real-world datasets. Moreover, due to the Markov property, Markov models does not memorize any past stages, while past stages can still matter or even alter outcome of generation. Such property also leads to the limitations in learning high-dimensional and complex data distributions [4].

Naive Bayes is a probabilistic generative model fundamentally based on Bayes' theorem, which computed probability of an event based on prior knowledge of conditions on related events. Naive Bayes learns the probabilities of each feature conditioned on each class label by analyzing and estimating likelihood estimations of these probabilities in the given training data, and then they aim to find a class label that maximizes the probability of that class given the data features. Techniques of modeling Naive Bayes usually involves maximum likelihood estimation or Bayesian methods.

Naive Bayes uses the assumption that all data features are independent, given their class label. Such assumptions allows the simplicity and efficiency for calculation for Naive Bayes, especially in large high-dimensional datasets, and thus broadly used some classification tasks. However, since Naive Bayes assumes the data are conditionally independent, they show limitations in modeling data with correlations [4].

2.1.2 Variational Autoencoders

Variational autoencoders (VAEs), proposed by Kingma and Welling in 2013 [34], offer an alternative approach to model generative models and gain their reputation by its promising ability to generate high-quality synthetic data samples.

VAEs are based on autoencoders, where a typical autoencoder consisits of two neural networks: an encoder and a decoder. The encoder compress input data to a smaller vector space, and the decoder attempts to revert the compressed vector to its original representation. Ideally, an autoencoder trains to learn to compress and depress data information with subtle loss. However, the vector spaces that a standard autoencoder compressed into do not secure continuation, which repents a problem are generative models where drawing random samples from a continuous latent space is desired.

VAEs, as a special kind of autoencoder, resolves the problem of latent space continuation by enforcing a mapping on input data features to a probabilistic latent space that corresponds to the parameters of a variational distribution. VAEs typically fulfill this by optimizing a proabilistic lower bound on the log-likelihood of the data, making them more interpretable and providing a principled framework for learning latent representations.

VAEs show their remarkable ability in generating diverse data, no matter continuous or discrete, labeled or unlabelled, across multiple domains. However, VAEs remain one major limitation in the blurry of the data sample thet generated [9], caused by the loss from the process of compressing and recovering data distributions.

2.1.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs), proposed by Goodfellow et al. in 2014 [20], poss another powerful approach for generating synthetic data. The structure of GANs consists of two neural networks: a generator and a discriminator. The generator aims to generate data samples that attempts to resemble distributions of training data, while the discriminator aims to distinguish between synthetic and real data samples. During the training process, the generator and the discriminator compete against an adversary by performing a minimax trick where the generator attempts to minimize the difference between the distributions of synthetic and real data samples and the discriminator attempts to maximize its ability to differentiate between synthetic and real data samples. This adversary in the training process enables both the generator and discriminator learns to be robust. After the training process, the discriminator is discarded and generator is kept to generate data samples.

GANs are able to closely resemble the distributions of real data and generate high-quality realistic data samples, gaining broad attention and broad uses across multiple domains, such as computer vision and natural language processing. However, GAN models still represent limitations. The discriminator in GAN only learns to differentiate between real and synthetic data. It does not learn features of the data samples, resulting possibly synthetic data samples with similar style compared to real data but not what actually exists.

2.1.4 Flow-based Generative Models

Flow-based generative models [35, 48] are generative models that model data distributions of given data utilizing the technique of normalizing flow, which targets to find invertible transformations to map a simple base distribution, typically a Gaussian or uniform distribution, to the more complex target data distribution [57]. Since the transformations are enforced to be invertible, it enables both data generation and likelihood estimation.

Flow-based models shows great ability to generate high-quality synthetic data samples. Flow-based generative models find applications across multiple domains, including text generation, image generation, and speech generation. Unlike traditional generative models such as Variational Autoencoders (VAEs) [34] or Generative Adversarial Networks (GANs) [20], since flow-based models offer exact likelihood estimation, it enables better assessment of uncertainty of the models and thus more stable results over the generated samples [7].

2.1.5 Generative Pre-Trained Transformer

Generative Pre-trained Transformers (GPTs) [53] represent a groundbreaking advancement in natural language processing (NLP) and generative modeling. GPT models show remarkably human-like ability in text generation. Especially in text prediction based on given preceding text and showing, GPT models poss incredible ability to understand and produce coherent and contextually relevant text based on given inputs of text sequences. GPT models have been utilized across multiple domain for text related tasks, such as translation and question answering.

GPT-based models are build on transformer models [64]. GPT models typically employ a decoder-only transformer setup with multi-head self-attention layers augmented with positional encodings to capture the sequential characteristics of input data and residual connection among normalized layers. Then, GPT models gain its power in generation tasks through pre-training on vast amounts of text data, enabling the model to learn patterns and structures of language. For example, GPT-4 (one of the most notable iterations of GPT models) has an unprecedented scale with 175 billion parameters. During fine-tuning or inference, GPT models autoregressively predict the next token based on the given data sequence.

Beyond text generation, GPT-based models showed to be effective to be adapted into other generation tasks, especially image generation [72, 75]. One example is the DALL-E model [56], which treats pixels as sequential data and is then able to generate high quality images from given text prompts, demonstrating the potential of GPT-based architectures beyond tasks related to natural language.

2.1.6 Latent Diffusion

One of the most powerful image generation model nowadays is the latent diffusion models [58], trained on large-scale, multi-modal dataset LAION-5B [60]. They have the ability to

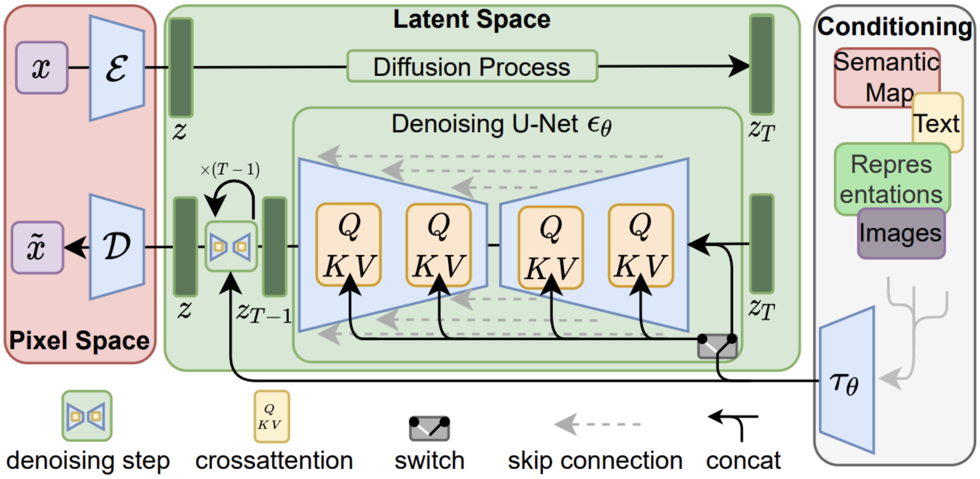


Figure 2.1: Architecture of latent diffusion models. Reproduced from [58].

generate high-quality, high-resolution, realistic image data samples, and thus is broadly utilized in multiple domains.

Diffusion models are characterized by the diffusion process, which can be split into a forward and a backward diffusion process. The forward diffusion process attempts to deconstruct the structure of the data distribution by gradually adding random noise to the data distribution based on a certain schedule. After the forward diffusion process, we should be able to get mimetic pure random noise. The reverse diffusion process, on the other hand, attempts to restore the structure of the data distribution through gradually denoising the random noise in a given timestep, which is shown to be efficient and relatively easy to train [63]. The denoising model utilized a time-conditioned U-Net, augmented with cross-attention mechanism conditioned on information of the data. Symbolically, for each input y, the cross-attention blocks are conditioned on

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d}}) \cdot V,$$

where $Q = W_Q^{(i)} \cdot \varphi_i(z^t)$, $K = W_K^{(i)} \cdot \tau_\theta(y)$, $V = W_V^{(i)} \cdot \tau_\theta(y)$, and $\varphi_i(z_t) \in \mathbb{R}^{N\times d_\epsilon^i}$ denotes a intermediate representation of the UNet implementing ϵ_θ, and finally $W_V^{(i)} \in \mathbb{R}^{d\times d_\epsilon^i}$, $W_Q^{(i)} \in \mathbb{R}^{d\times d_r}$ and $W_K^{(i)} \in \mathbb{R}^{d\times d_r}$ are three learnable projection matrices.
In Figure 2.1, we show the full diagram for architecture of latent diffusion models [58].

Through training through the diffusion process, latent diffusion models gain the ability to generate a variety of high-resolution synthetic image datasets, showing their tractability and flexibility. They also present better quality synthetic data samples compared to VAEs and GANs, due to its semantic compression which performs less information loss during the training process. Diffusion models represents an upper bound of current generative models, and thus, for the purpose of this thesis, we focused on research latent diffusion models as a focus for generative models.

Even though diffusion models shows impressive ability in generating synthetic images, they still presents challenge in the data and computational resources required for training.

2.1.7 Restrictions

Though different generative model have shown promising future for the task of generating synthetic data, limitations among different preserves. Moreover, another important restriction of the current generated models is the domain of data which the models can generated. The ability of high-quality generation for those models is strictly restricted by the domain of the large-scale input datasets that are used to train those models [76]. However, the task and ability of out-of-domain generation task is also important, which we will discuss in more detail in section 2.3.

2.2 Data Efficiency

Though generative models have shown excellent generated results in text, images, videos and etc., current generative models still present challenges and limitations in the data and computational resources that they required to train on. For example, the powerful latent diffusion model [58] (discussed in §2.1.6) is trained on the LAION-5B dataset [60], which contains around 170M high-quality and high resolution data samples. Such large-scale datasets is almost impossible in multiple domains. Medical domain, in particular, might be insufficient of 170M patient samples for some relatively rare disease diagnoses. Thus, there is an rising importance in increase the efficiency in the usage of the limited data during the training process [29]. There are several approaches to improve models' data efficiency. We will discuss two approaches in particular related to our thesis: a cluster-based approach and active learning approach.

2.2.1 Cluster-based Approach

One way to improve data efficiency is cluster-based methods. Cluster-based methods attempt to capture relationships and structure of data features, and then group "similar" data samples together into clusters. In this way, cluster-based methods facilitate certain data compression based on data features. Models take advantage of this by learning clusters instead of individual data samples, and are able to achieve similar performance but trained

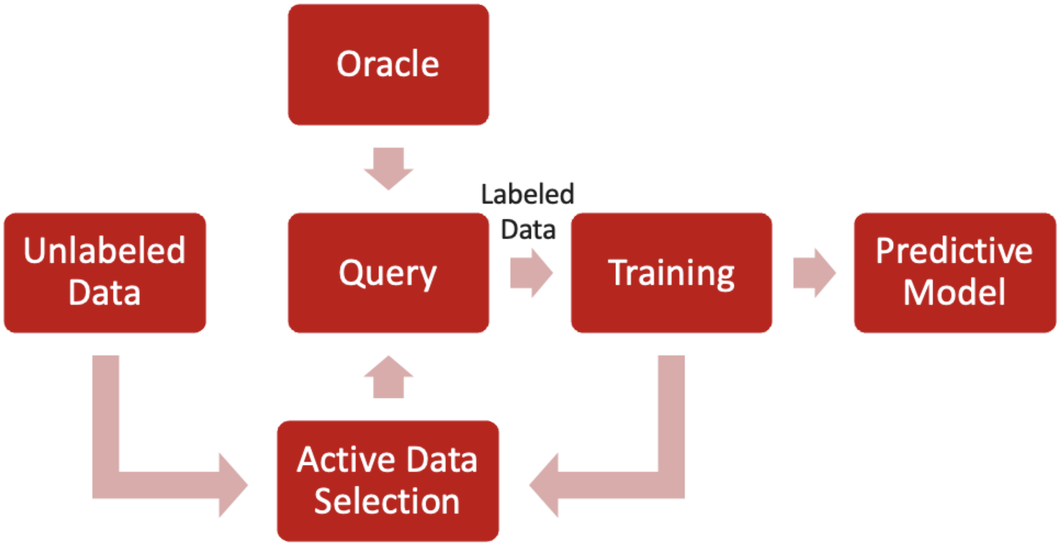


Figure 2.2: Three main active learning approaches. Reproduced from [61].

on less data input. Such process largely simplified the learning task and enables a significant decrease in computational resources required for training, and also improving the data efficiency for the overall model, especially shows to be efficient for dealing with large-scale datasets. It also mitigates issues like overfitting and improves the generalization ability of machine learning models. Such methods have already be shown to be efficient in improving model efficiency among multiple machine learning models [19, 13].

2.2.2 Active-learning-based Approach

Another approach to improve data efficiency during training is though active learning methods. They selectively select data samples to train on and to modify model parameters instead of taking consideration of each data sample in a large-scale dataset. The fundamental of active learning in machine learning is to train a model that can actively query an oracle to label the most informative instances, thus saving training effort while maximizing the model's performance. Figure 2.2 shows the general process of active learning. The idea is inspired my real-world active learning in classrooms, where students actively get feedback from teachers to improve their learning outcomes.

The core algorithmic idea of active learning is to label and learn data samples nearer best decision boundary the model currently learned. Figure 2.3 illustrate a diagram for the general idea of active learning.

There are three major approaches for active learning: 1) pool-based active learning, which

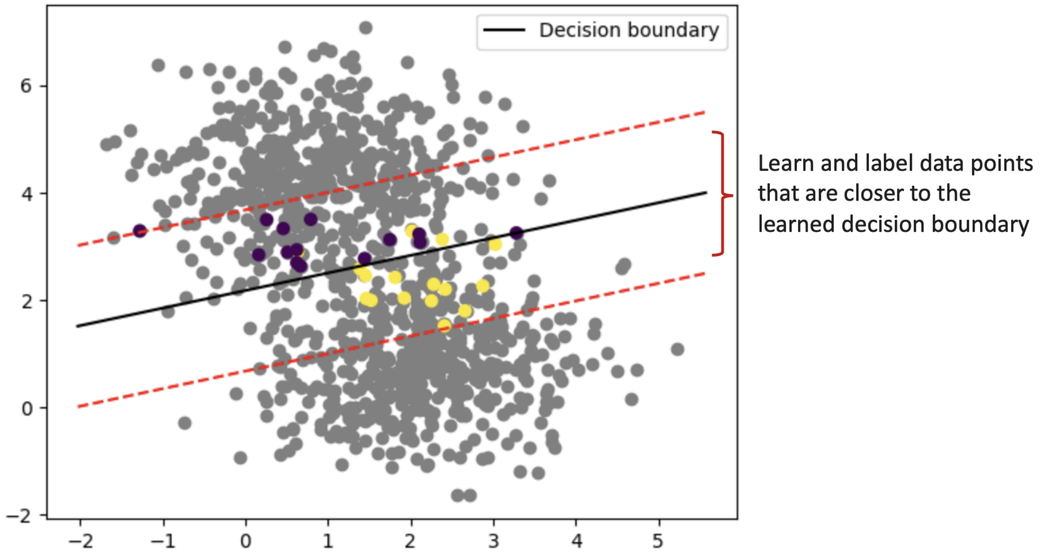


Figure 2.3: General idea for active learning.

the model has the query orable to tell the information usefulness of all data in the sample pool and greedily select the top N samples that are most informative; 2) stream-based selective sampling, which is similar to pool-based active learning but the model also decides whether to query for the ground truth label depend on its confidence threshold of the predicted label; 3) membership query synthesis, which models query on synthetic samples it generates [61]. Figure 2.4 depicts the methods and differences among these three approaches.

Moreover, active learning methods combined with clustering-based techniques also shows improving the performance of multiple models across different tasks [47, 5], which is also a major approach this thesis will be taking. We will discuss our methods more in detail in Chapter 3.

2.3 Out-of-domain Generation

The generated results of generative models (such as latent diffusion [58]) also has restrictions based on the source domain of their training set, which is a typical limitation of current machine learning models [76, 45]. As mentioned above in section 2.1, diffusion models [58] are trained on LAION-5B dataset [60], which is a large-scale dataset but mainly contains image of general purposes, indicating that the ability of generating images out of the domain of LAION-5B can be limited. Recent works has shown the failure of original latent diffusion model in generating out-of-domain medical health images [11], while out-of-domain generation can be essentially helpful. As mentioned in Chapter 1 (section 1.1.1), synthetic

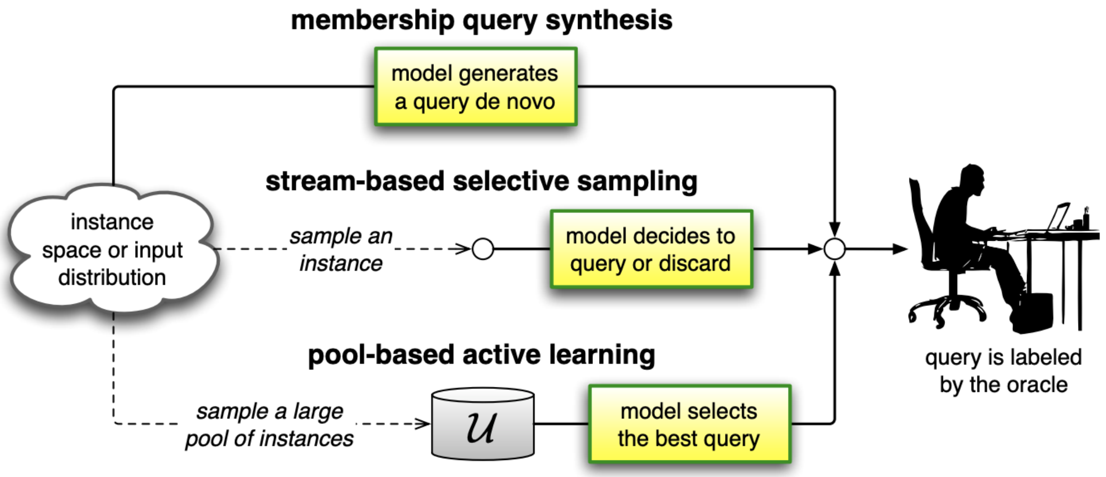

membership query synthesis
model generates a query de novo
stream-based selective sampling
instance space or input distribution
sample an instance
model decides to query or discard
pool-based active learning
sample a large pool of instances
$\mathcal{U}$
model selects the best query
query is labeled by the oracle

Figure 2.4: Three main active learning approaches. Reproduced from [61].

high-fidelity datasets have promising future use to augment machine learning models across multiple domain and serve as a possible solution for the data paucity problem. The more out-of-domain generation task the models can overcome, the more we can adapt machine learning models into different domain. With the expectation upon the generalization of machine learning models to enhance human ability across multiple various domain, there exists an increasing need to overcome the out-of-domain generation tasks.

2.3.1 Domain Adaption

One popular approach to overcome the challenges of out-of-domain generation tasks is the domain adaption technique. It aims to transfer knowledge learned from a source domain to one targeted not-leaned domain with the purpose to improve performance in the out-of-domain tasks, helping to bridge the gap between the source domain and the targeted domain [3, 18]. There are several approaches for domain adaption, which focus varies from data to feature to model [74, 18].

Data-based Adaption is specifically concerned with the difference in distributions learned due to difference in sampled data instances (e.g. due to sample selection bias [28]), and attempts to align individual data instance into the targeted domain. Efficient methods of implementing this can be data reweighting in the source domain to fit the targeted data space [74, 69].

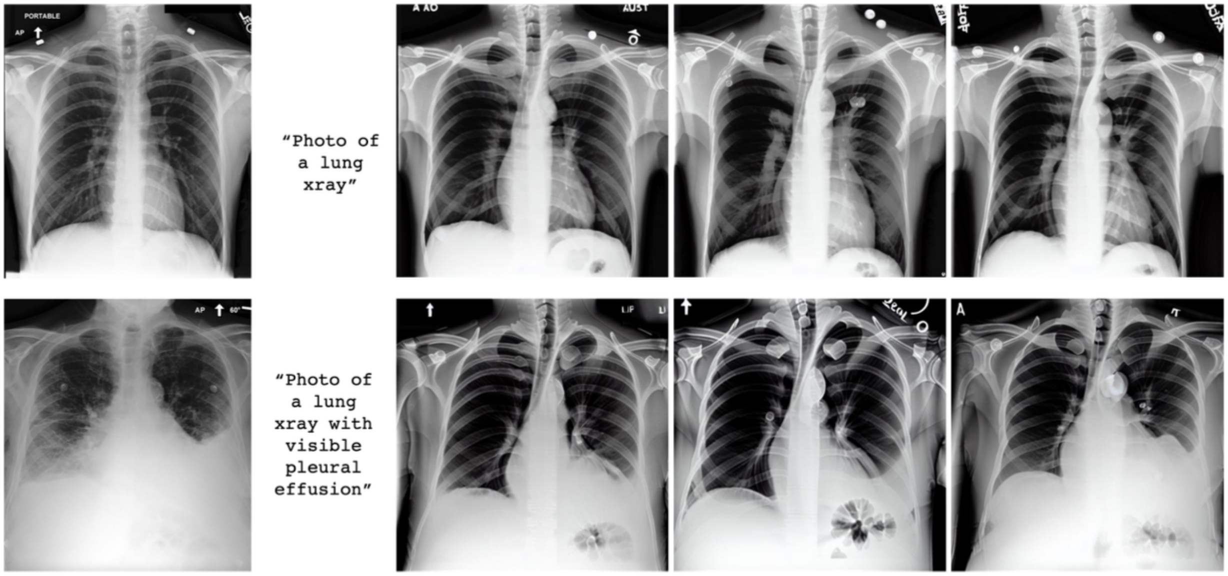


Figure 2.5: Some examples of domain specific synthetic images for adapting diffusion models for specific downstream medical image generation task. Reproduced from [11].

Feature-based Adaption learns a transformation between the feature spaces of the source domain and the targeted domain, and has shown to be efficient in transfer learning to targeted domains [74, 41, 39, 68].

Model-based Adaption directly adapt models' parameters learned based on source domain to fit the targeted domain. A commonly used technique to implement this is through model parameter fine-tuning, which usually implemented by neural networks [74, 42], and showed powerful results in adapting different domains for generative tasks [11]. Figure 2.5 [11] shows some examples of adapting diffusion models for the downstream task of medical imaging (specifically frontal chest X-rays) generation. We will discuss more about models for fine-tuning in more detail in section 2.3.

All three listed approaches have shown promising results in the task of domain adaption. However, it still remains a question on how to effectively adapt the domains of models, as mentioned in section 2.2.

2.4 Model Fine-tuning

Fine-tuning is a commonly used technique to modify model weights to transfer learning to a specific domain. Pre-trained large image models tends to lack the ability to produce specific.

Fine-tuning in machine learning involves taking a pre-trained model and refining its parameters to fit a specific task or dataset. This process is particularly useful when dealing with limited data or when aiming to improve the performance of a model on a particular task. Initially, a pretrained model trained on a large-scale general dataset is selected. Then, the model architecture is modified to suit the requirements of the target task, which may involve adjusting layers, activation functions, or the output structure. During training, the parameters of the modified model are updated using a task-specific dataset, allowing the model to learn task-specific features while retaining the knowledge gained from the pre-training. Hyperparameters are fine-tuned to optimize performance, and the model's performance is evaluated on a validation set. Fine-tuning is an iterative process, often involving multiple training cycles and adjustments to achieve the best performance. Fine-tuning enables the transfer of knowledge from pre-trained models to specific tasks, resulting in improved efficiency and effectiveness in machine learning applications.

2.4.1 Dreambooth

Our thesis focused on the use of fine-tuning for adapting model to out-of-domain generation tasks. One popular model to implement this technique is the Dreambooth fine-tuning model, introduced by Ruiz et al. in 2023 [59].

The Dreambooth model enables fine-tuning large-scale text-to-image models to adapt domain of input dataset for some downstream tasks within fewer input samples, which is especially helpful for situations where data samples are limited.

The general pipeline of Dreambooth model functions as follows: The model takes in input image data of domain we targeted to fine tune on paired with their text prompts, and a text-to-image diffusion model. It then pairs a unique identifier with the text prompt and perform Dreambooth training. The training outputs an fine-tuned text-to-image model that encodes a unique identifier that refers to the tuning set. At the inference stage, we utilize the unique identifier to generate data in our targeted domain. Figure 2.6 shows a graph of how Dreambooth pipeline works.

During the training process, the models first take the given image-text pairs with the unique identifier and fine tune the low-resolution text-to-image model while applying a class-specific prior preservation loss, which allows the model to learn new data while preserving

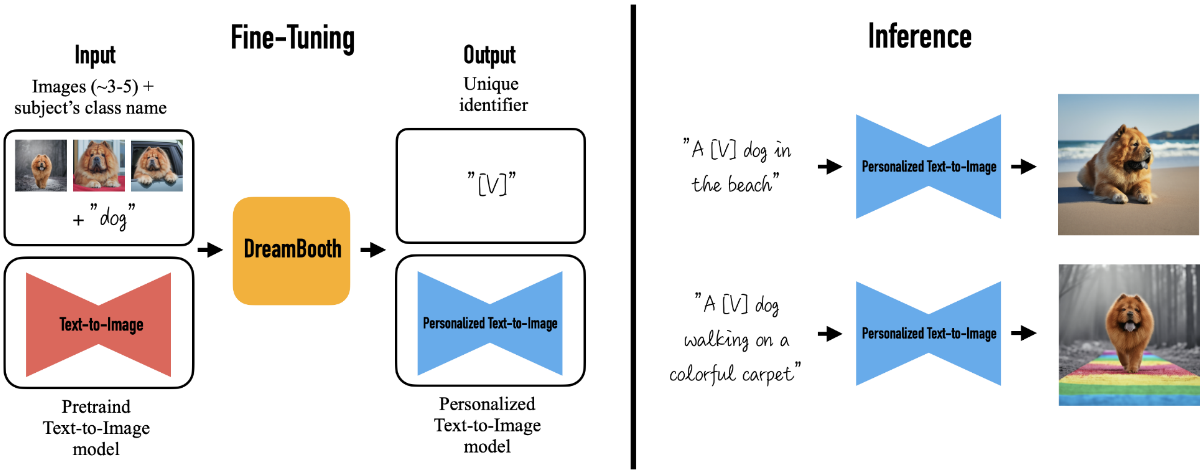

Fine-Tuning
Input
Images (~3-5) + subject's class name
+ "dog"
Text-to-Image
Pretraind Text-to-Image model
DreamBooth
Output
Unique identifier
"[V]"
Personalized Text-to-Image
Personalized Text-to-Image model
Inference
"A [V] dog in the beach"
Personalized Text-to-Image
"A [V] dog walking on a colorful carpet"
Personalized Text-to-Image

Figure 2.6: Pipeline for Dreambooth. Reproduced from [59].

pre-trained knowledge by utilizing the semantic prior that the model has. The model then fine-tune the super resolution components of the model with pairs of low-resolution and high-resolution images taken from our input images set. Figure 2.7 shows a graph of architecture of Dreambooth model.

Dreambooth model has been shown to be effective in adapting generative models and transfer model learning for specific downstream generation tasks. For exmaple, Chambon et al. show its effectiveness in adapting diffusion models into task of chest X-ray generation. Examples shown in figure 2.5.

2.4.2 Efficient Fine-tuning

As discussed in section 2.2, the efficiency of machine models remains a challenging but important task. So as model fine-tuning. Multiple research approaches have taken efforts to optimize the process of model adaption for efficient fine-tuning. Among all of the approaches, our thesis is interested in whether more general approaches also works for fine-tuning.(such as cluster-based methods and active learning as mentioned in section 2.2.1 and 2.2.2). We found cluster-based methods still efficient for increasing model efficiency under fine-tuning background [49, 73]. More specifically, weight initialition of data points or clusters can also alter the performance of efficiency of fine-tuning [16].

Also, active learning approaches are shown be an effective technique in improving fine-tuning efficiency [2, 71]. Specially, Bashar and Nayak proposed method of setting class distance boundary and utilize the boundary threshold for active data selection during the

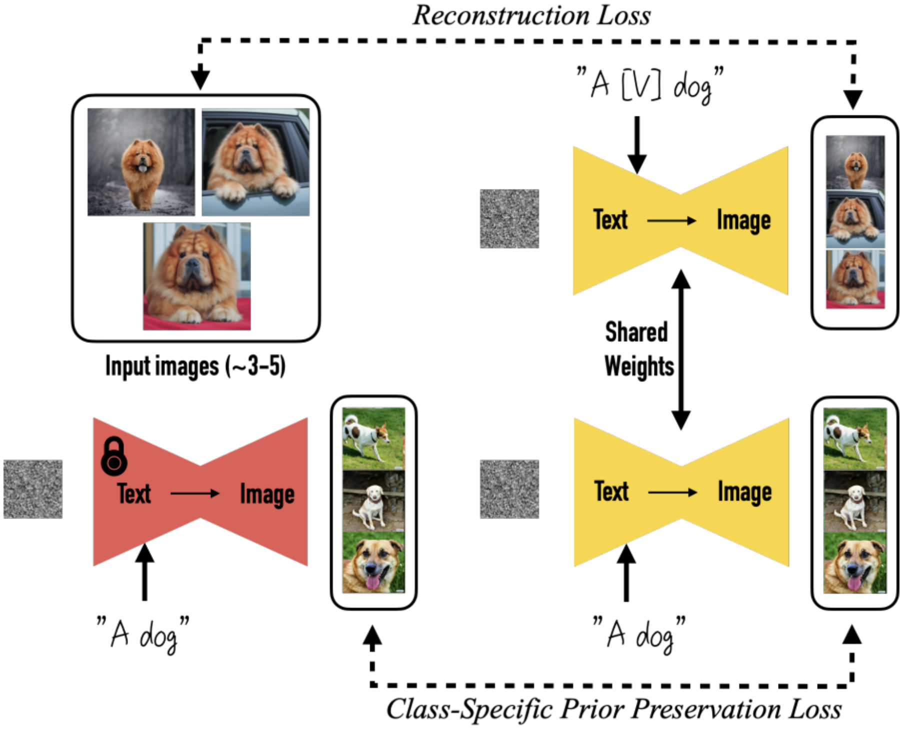


Figure 2.7: Model architecture for Dreambooth. Reproduced from [59].

fine-tuning process. This method has been shown to efficiently fine-tuning transfer learning for downstream tasks [2]. Figure 2.8 shows the pipeline for an actvie-learned embedded fine-tuning approach discussed in Bashar and Nayak's paper [2]. Our experiments get inspirations from this architecture, which we will discuss in more detail in Chapter 3 and Chapter 4.

2.5 Open Challenges

As discussed in Chapter 1, our thesis is interested in investigating methods for out-of-domain generation. In this chapter, we discussed different approaches in resolving out-of-domain generation challenge. We found that model fine-tuning shows convincing power in adapting domain-specific generation tasks and also present less limitations on data domains. Also, current research seems to lack of discussing on how well synthetic data generated for model after domain adaption techniques can be used for training purposes, while this task present significant solutions to the data limitation or paucity in specific domains.

Moreover, fine-tuning large-scale models can still require massive scale tuning set, and the data paucity problem in some domains can still block the way. The data efficiency problem still requires concerns and researches.

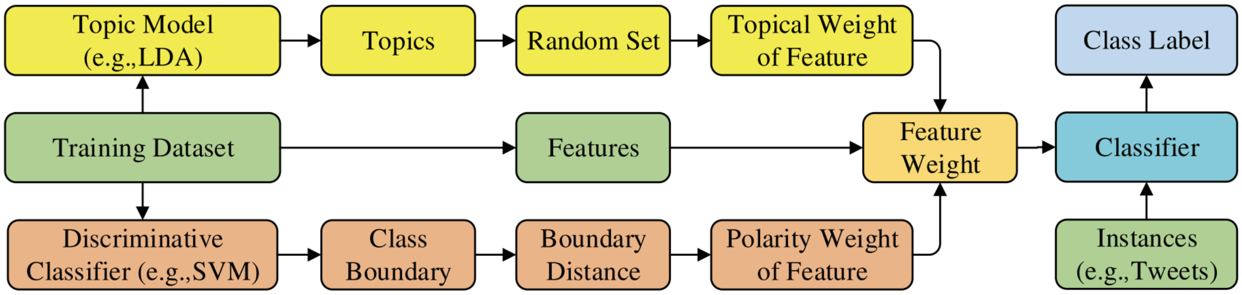


Figure 2.8: Model pipeline for active-learning-based fine-tuning. Reproduced from [2].

These collectively form the research question of thesis: How to evaluate how well a synthetic dataset can be trained i.e. the trainability performance of a synthetic dataset, and how to efficiently fine-tuning models for domain adaption for out-of-domain generation tasks given fair trainability performance of the synthetic datasets. After confirming our research question, we then proposed our research approaches and methods based on methods we found effective in our literature review, which we will discuss in detail in next chapter (Chapter 3).

Chapter 3

Methods and Approaches

The problem and the main research purpose outline in Chapter 1 together described the goal of this thesis: investigate methods for domain-specific/out-of-domain generation for generative models, mostly focusing on better evaluated performance for model training. In this chapter, we will discuss the research approaches and methods we used to investigate the problem.

3.1 Domain Focus: Frontal Chest X-rays

As we are mainly interested in understanding efficient methods for model adaption for out-of-domain generation tasks. Deploying machine learning models in medical domains shows especially beneficial [62, 44], while current generation models does not seem to show promising ability in generating medical domain-specific data, especially in medical imaging domain [11]. For example, the left figure of figure 3.1 presents an example of a synthetic frontal chest x-ray generated by latent Diffusion model [58] without any domain adaption technique. We see the model does not reproduce the exact features of realistic frontal chest X-rays. However, when applying domain adaption techniques, such as model fine-tuning, the model is able to generate realistic-looking frontal chest x-rays (see the right figure of figure 3.1). Based on these facts, we foresee the application of generative model in medical domain, specially in medical imaging on frontal chest x-rays.

Moreover, based on literature review (Chapter 2) on related work in domain adaption in medical imagining, we see previous researches showed promising results in using fine-tuning model Dreambooth to adapt generative image model into frontal chest X-ray generation tasks [11, 12]. Considering a promising future of deploying generative model in medical domain-specific dataset generation tasks and also to get insights of our methods compared other experimental results from previous researches, this thesis focused on a the specific

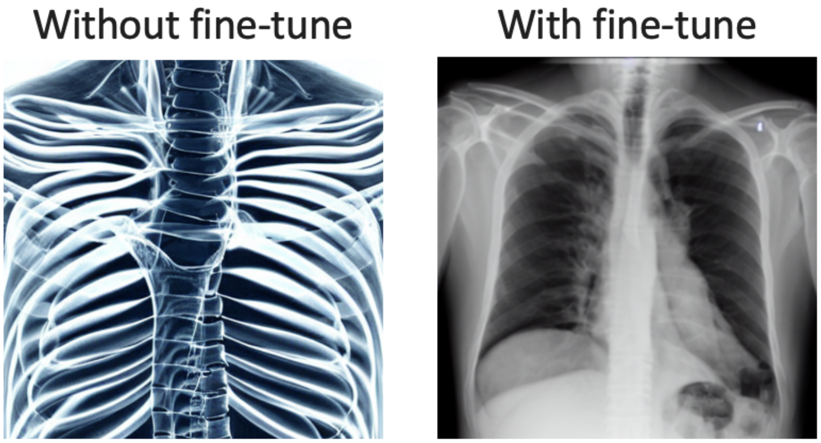


Figure 3.1: Comparision of the result of chest X-ray images we generated: without fine-tune (left) and with fine-tune (right).

domain of medical imaging generative of frontal chest X-rays.

3.2 Research Tools

As breifly discussed in Chapter 1, the thesis is interested in synthetic data generation as a possible solution for the data paucity problem across different domains. It then becomes significant to measure how well a sythetic dataset are able to achieve similar training performance compared to real datasets, which, in this thesis, refers to the trainability of dataset (i.e. how the dataset can train a model). Yet, in our literature review and to the author's knowledge, not much research touched on measurements in the trainability of synthetic datasets. In this thesis, we proposed a new evaluation pipeline along with a new measureing metric to better evaluation the task the thesis focused to investigate on.

3.2.1 2-way Evaluation Pipeline

In our literature review, we found that most current research tested the quality of synthetic dataset by training a classifier model, which has been proved to have good performance in the classification task, to classify whether the synthetic data sample belong to the class it was tasked to generate. They argue that the task can tell is synthetic images are able to reproduce features of the targeted class [11]. However, such evaluation task cannot avoid the generative model simply overfitting the input data and simply reproducing features of input images. And, as mentioned in Chapter 1, if we see the task of synthetic data generation as a solution to the data paucity problem in training machine learning models, we need to see how well the performance of models trained on synthetic datasets can be.

Therefore, we proposed a 2-way evaluation pipeline for measuring the task of synthetic dataset generation:

1. Train a classifier model using synthetic dataset, and then test the trained classifier using real-world dataset.

2. Train a classifier model using real-world dataset, and then test the trained classifier using synthetic dataset.

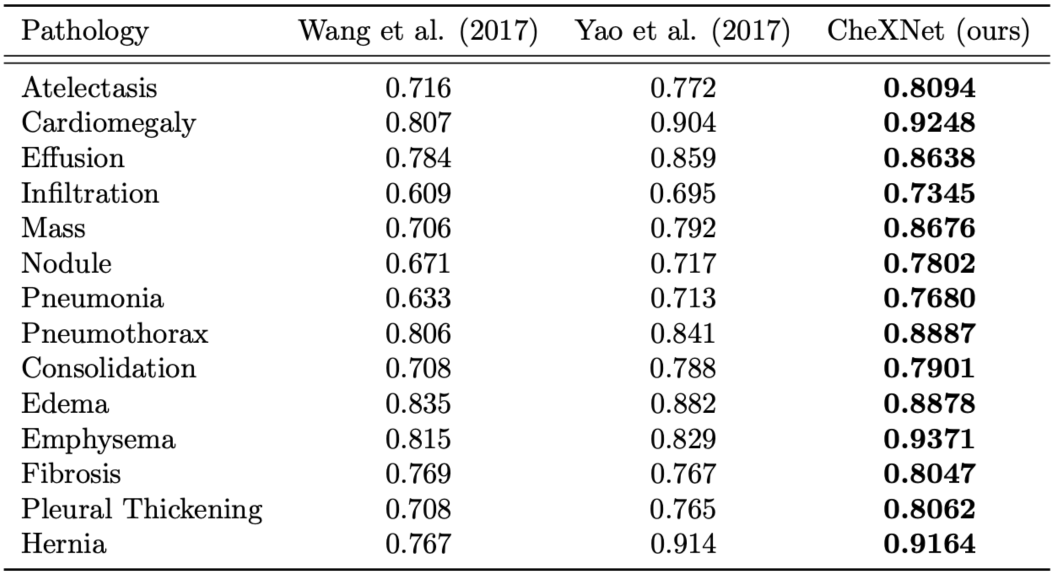
We then compare the resulting performance of the two classifier from the two pipes.

Figure 1.1 presents the diagram of our evaluation pipeline. In this 2-way evaluation pipeline, we are able to see and compare how well the synthetic images can be trained on, while preserving the classical approach of observing how accurate the synthetic features of the data samples performs. Good synthetic datasets should be able to reproduce the distribution of real-world dataset close enough to perform relatively and similarly well in both ways.

Classification Model: CheXNet

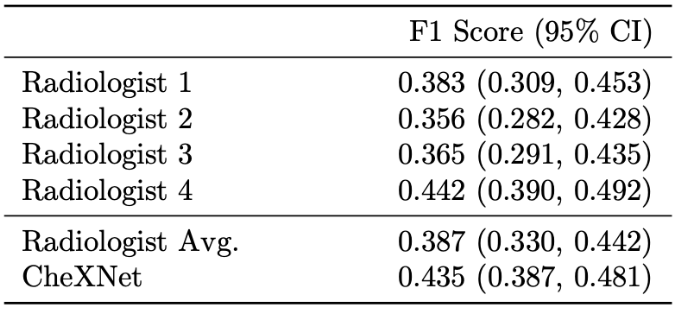
As mentioned above, our evaluation pipeline depends on the analysis of the compared performance of the two classification models, one trained on synthetic images and one trained on real-world images, to gain insights on the quality of synthetic dataset. The validation of such analysis fundamentally based on the assumption on the robustness of architecture of the classification model that the classification model is able to learn and classify well given high-quality data. Therefore, deploying a robust classification model for the domain-specific downstream classification in evaluation pipeline is extremely important.

Considering the domain of frontal Chest X-rays, our thesis deployed the CheXNet classification model [55]. CheXNet is classification model architecture specifically designed for the task of differentiating 15 classes of frontal Chest X-rays, and have shown a high performance in such task when trained on a high quality Chest X-ray dataset, NIH ChestX-ray14 dataset [70] (details on AUROC score presented in figure 3.2). CheXNet even shows better performance in F1 score for some classification tasks than the average of radiologist (details show in figure 3.3) [55]. Such high performance in classification tasks indicates the robustness of the model architecture, and thus is a suitable model to deploy in our evaluation pipeline. For the purpose of this project, we simply the CheXNet model to binary tasks which only differentiate between health and unhealthy frontal chest X-rays, and deploy the model into our evaluation pipeline.



Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (ours)
Atelectasis	0.716	0.772	**0.8094**
Cardiomegaly	0.807	0.904	**0.9248**
Effusion	0.784	0.859	**0.8638**
Infiltration	0.609	0.695	**0.7345**
Mass	0.706	0.792	**0.8676**
Nodule	0.671	0.717	**0.7802**
Pneumonia	0.633	0.713	**0.7680**
Pneumothorax	0.806	0.841	**0.8887**
Consolidation	0.708	0.788	**0.7901**
Edema	0.835	0.882	**0.8878**
Emphysema	0.815	0.829	**0.9371**
Fibrosis	0.769	0.767	**0.8047**
Pleural Thickening	0.708	0.765	**0.8062**
Hernia	0.767	0.914	**0.9164**

Figure 3.2: Performance of CheXNet: AUROC. Reproduced from [55].



	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Figure 3.3: Performance of CheXNet: F1 score compared to radiologist. Reproduced from [55].

3.2.2 Metric: Fine-tune Efficiency Ratio

As mentioned in the above section, in our 2-way pipeline, we compare the resulting performance of the two classifier (one trained on synthetic data and test on real-world data, while the other trained on real-world data and test on synthetic data). Aside from accuracy measures of the classifiers, to better compare the numerical results of the performance of the two classificers. We proposed a new metric Fine-tuning Efficiency Ratio (FTER).

$$FTER = \frac{\text{AUROC of classification task trained on synthetic data test of real-world data}}{\text{AUROC of classification task trained on real-world data test on synthetic data}}$$

FTER is the ratio of AUROC score of the two trained classifiers, where AUROC tells how a trained model can make correct prediction on the classes [8]. High AUROC score and accuracy for a classification modes usually means that the model learns well in the classification task. FTER compares how well the two classifiers perform. FTER can fell into three cases:

1. FTER = 1. The two classifier make predictions perfectly similar. The synthetic data perfectly learns the distribution of the real-world data.

2. FTER > 1. The classifier trained on synthetic images predicts better. This means that the synthetic data not only capture the distributions of the real-world data. It also learns a better distribution for dataset to trained on such tasks than our pre-existing real-world dataset.

3. FTER $<$ 1. The classifier trained on synthetic images predicts worse than classifier trained on real-world data. The generative model might not well capture and reproduce the distribution of real-world data. If the FTER is specifically low, it might indicate that the model generates synthetic dataset not presenting full distribution of the real-world data, and thus possibly overfitting input training sets.

For targeting good generative model for the task of synthetic dataset generation for model training purposes, ideally we want models that have high accuracy and FTER score as close to 1 as possible or even FTER ≥ 1.

For our experiments, we ran our 2-way pipeline on different experiment scenarios, and, for each experiments, we compute the FTER score. We aim to find methods for efficient domain-specific fine-tuning with high accuracy and high FTER score.

3.3 Research Approaches

Based on our literature review, we concluded research approaches resolving related model efficiency problems into two main categories:

1. Targeting on data input, which focuses on data processing (e.g. clustering-based methods) to improve the quality of overall training data to achieve better performance.

2. Targeting on improving model data acquisition to improve efficiency for the training process. For example, the active learning approach during model training aim to learn more efficiency of the data.

Figure 3.4 shows the graph of research approaches we mainly focus on.

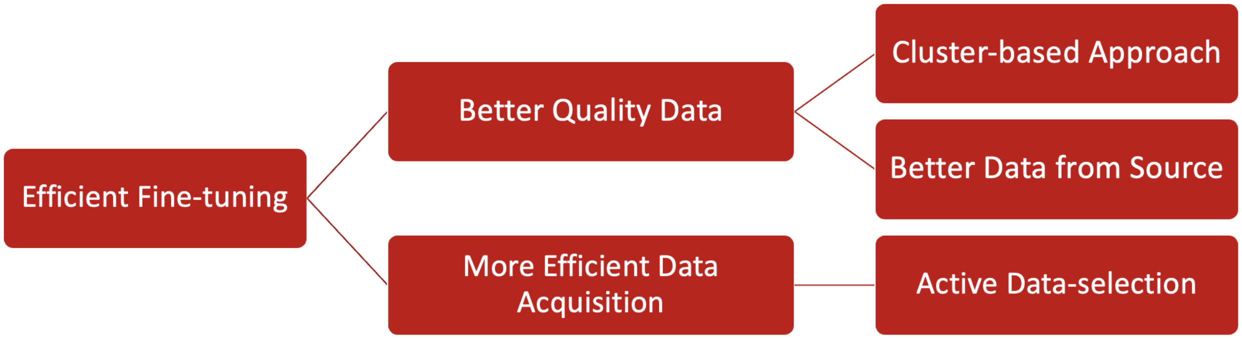


Figure 3.4: Graph of our research approaches.

Though previous research does not agree with our focus, we gain insights on research approaches we can take to investigate on improving fine-tuning efficiency for adapting out-of-domain generation tasks. Thus, we take these two approaches to explore methods for resolving our task. As discussed in section 2.5, we found that model fine-tuning shows to an efficient method for adapting models for out-of-domain generation tasks. Among models for fine-tuning, Dreambooth models stands out due to its proved ability of generating high-quality datasets for specific domains (e.g. medical imaging [11]). Therefore, the thesis will focus on taking Dreambooth models as a way for out-of-domain adaption tasks, and investigate on how different methods under the two major approaches performs.

3.3.1 Aiming for Better Quality Data

Quality and quantity of the training data appears to have great impact on model performance [43, 10]. The concept of "garbage in, garbage out" underscores the importance of high quality datasets taken as inputs for model training. Generally, there are two approaches to make the training data to have higher quality. One is to collect better data from source, which takes effort in manually identify high-quality data samples. The other is using algorithm techniques to select high-quality data samples from a collected dataset and deploy this higher-quality subset of data into training. One popular technique in this approach is clustering, which is also one method this thesis will focus on.

Better Data from Source

One possible approach to take to aim for better quality data is to fine better quality dataset from its source, where data quality in machine learning usually refers to how well informative a data sample can be to the models [22]. However, it remains a question on what kind of data samples are more informative and easier to learn.

Machine learning is often analogized to the task of educating a model. Then consider the situation of real-world education. When students want to learn something, what might be some of the more reliable and informative physical sources? One answer is simply textbooks. Textbooks are made of selected information by professional experts who think these information can be more educational. Research also shows model training on textbook-based datasets shows better performance on language models for code [23]. Now we narrow our sight a bit into the image domain. For the images in textbooks, they are chosen possibly because professional experts consider those image datasets as more useful, typical and characteristic for educational purposes, which basically means more informative data samples. Thus, if we collect image data samples from textbooks, we are able to form a high-quality dataset.

In fact, during our research, we compared data samples in our collected textbook dataset and the pre-existing Chest X-ray datasets (e.g. ChestX-ray14 [70]). We found that pre-existing dataset do have non-typical medical cases, which can be hard to learn for both human and models.

Therefore, we manually collected over 1,000 images from medical Chest X-ray textbooks and form our textbook dataset. We performed our pipeline and results on this dataset and compared results. More experimental details will be presented in Chapter 4.

Cluster-based Data Selection for Training Data

Another possible approach to obtain better quality data is through clustering. Cluster-based data selection involves presenting data samples in some vector space and then grouping similar data points together into clusters based on features represented in the vector space. Then, the model selects training samples from each cluster. In this way, models are able to reduce the redundancy within the dataset and learn qualitatively less but more representative data samples. Since cluster-based data selection methods only learn selected number of data samples from each cluster (usually not the full dataset), models can significantly learn less data points, saving computational and data resources and improving model efficiency. Through balancing the number of selection selected from each cluster, models are able to resolve possible imbalances of classes in the training set, avoiding over-biased model outcome. Specifically, we deploy clustering models to cluster similar data points and pick samples from each data point. Depends on the number of samples from each cluster, we manually attempt to re-weight each cluster. Finally, we group the data together and put them into the stage of fine-tuning. Figure 3.5 shows the diagram for the process of our cluster-based data selection method.

There are usually 2 ways to find clusters. One is to cluster data based on their real-world

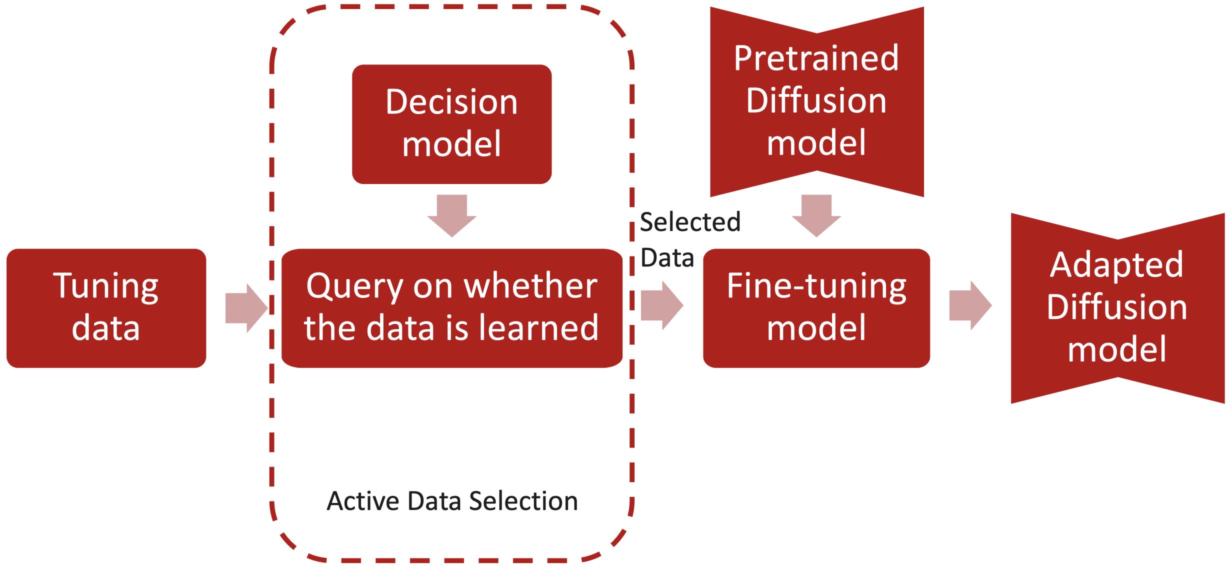


Figure 3.5: Diagram for the process of cluster-based data selection

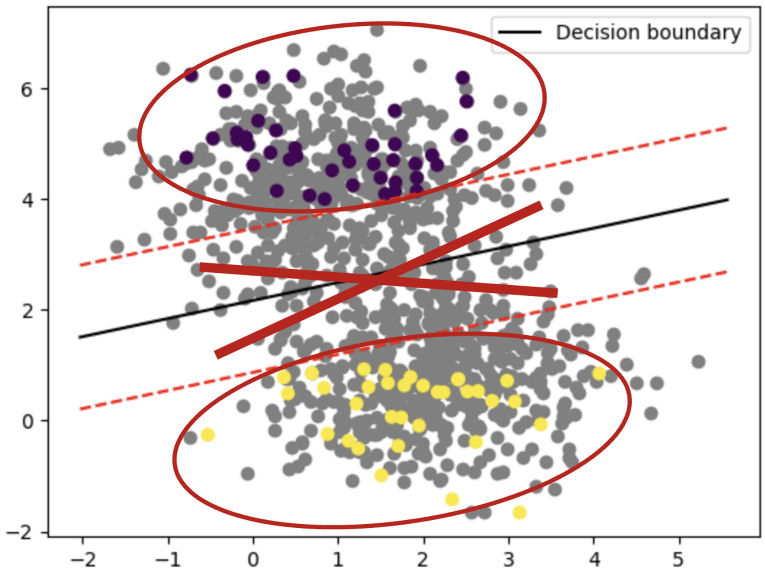
sub-classes. In machine learning, this usually requires additional information on the dataset. The other way is clustering algorithms, includes K-means algorithms and spectral clustering algorithm. This thesis is curious in how these two approaches differ in our task settings, and experimented on both approaches. Experiment details are represented in Chapter 4.

3.3.2 Aiming for More Efficient Data Acquisition

Optimizing the data acquisition for models is one important approach in improving efficiency and performance of machine learning models. More suitable data acquisition techniques for downstream tasks usually secure better training performance. There are various methods to improve the efficiency of data acquisition. This thesis focuses on active-data-selection-based approaches.

Active Data-selection Embedded in Fine-tuning Model

As discussed in section 2.2.2, active-learning-based approach have been shown to improve data efficiency for machine learning models. Active-learning based data selection methods involve filtering most important data points based on annotation from third-party oracle, and then only learn those important data points. This approach enhances model performance by learning more important and less noisy data, and also improve model efficiency by training on less data samples.



Only learn the important ones that are outside the boundary (not confident to generate)

Discard confident data samples

Figure 3.6: Idea of our active learning for fine-tuning for generative model

Extending this principle to the task of domain adaptation for out-of-domain synthetic data generation, we envision the application of active data selection among fine-tuning models to improve the models' performance and efficiency. Traditionally, active-learning techniques involve learning data closer to decision boundary so that the model can only learn more important data points which the model might not be exactly confident on to shape a more precise decision boundary. Under background of generative models, active learning techniques take an opposite direction. Generative models actively learn on data samples with distributions that are mostly unseen and unpredictable. That is generative models learns data samples with the greatest variance compared to its learn knowledge of data distributions so that the model is able to learn to generate with more variety. Figure 3.6 shows a disgram of the general idea of active learning for fine-tuning for generative models.

Similar approach would work for the task of domain adaption for generative models. We want to actively tune with the data samples that can provide the greatest variance to alter the pre-trained domain and adapt to the targeted ones. In practice, our thesis focused on an shown-effective domain-adaption technique of Dreambooth fine-tuning model [59]. We embedded an active data selection layer to the Dreambooth fine-tuning model, enabling the model to learn most various data samples. By deploying active data selection in Dreambooth model, we aim to enhance the model's ability to learn and to reproduce more diverse data samples and then enhance its ability to adapt to diverse domains. Through this combination of active data selection technique, we anticipate significant advancements in trainability of synthetic datasets and model training efficiency. Figure 3.7 shows a diagram of the idea of

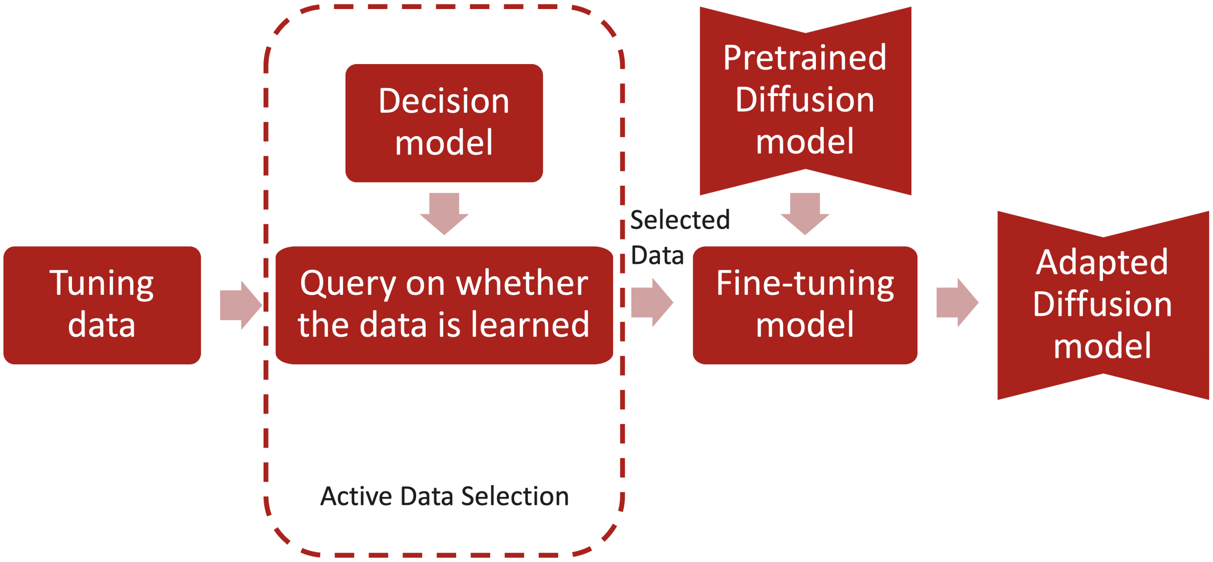


Figure 3.7: Diagram of our active data selection embedded fine-tuning model

our active data-selection embedded fine-tuning model.

Chapter 4

Experiments

Following presenting the methods deploys in this research, in this chapter, we will be presenting our research experiments in detail. For each experiment, we will introduce its experiment settings, then present the experiment results, and finally discuss the results.

The Blueprint Before diving into the experiment details of this research, we would like to depict a broad view of the blue print of our research experiments.

As introduced in chapter 1, the technique of synthetic datasets might serve as a future solution in solving the problem as data paucity for training massive models. In this blueprint, we foresee the synthetic dataset to have the following characteristics: 1) able to be easily generated; 2) able to represent correct information of the domain that the model is asked to generate.

Character 2, in particular, infers that, for any given task (e.g. classification, regression and etc.), synthetic datasets should be able to produce results similar to that of trained on real-world dataset. The idea forms a partial outline for the evaluation task that we are curious about: training some classifier model using synthetic dataset, and test using real-world dataset. For good-quality, well trainable synthetic datasets, the results should perform similar to a similar task but train on real-world data.

Researching the desired states can be hard, and more researches on characterizing the two characteristics are in urgent need. In this thesis, we take an approach on experimenting and measuring the two characteristics, and thus break down the task into the following research questions:

- How to evaluate trainability performance of synthetic dataset produced by fine-tuned generative model for out-of-domain generation.

- What are efficient methods of fine-tuning given relatively well trainability performance.

Keep the two research questions in mind, we shaped our first experiment of our initial thought on experimenting how well some random synthetic dataset can perform if it is deployed into training a classification task.

4.1 Experiment 1: Initial Approach

The general idea of our initial experiment is simple. We want to get some insights on how well synthetic dataset can performance when they are taken into training for some classification tasks. The simple thought is to train it. As discussed in 3.1, frontal Chest X-ray is one of the domains that are not yet learned by majority of pre-trained generative models and requires domain-adaption techniques, such as model fine-tuning to transfer learning into targeted domains. Therefore, we set our initial experiment as a simple task of testing domain adaption using fine-tuning technique to generate synthetic images, and then train and test on the synthetic images. The following sections will present the models we deployed, the results we got and what we observed.

4.1.1 Experiment Setting

In our initial though experiment,m we fine-tuned the image generation model, latent stable diffusion [58], using Dreambooth fine-tuning model [59] with randomly picked 1,000 images from a frontal Chest X-ray dataset, NIH ChestX-ray14 dataset [70]. Then, we generated 10,000 images with binary labels of either "Unhealthy" or "No finding" and put the synthetic images into training a chest X-ray classifier. We then task the trained classifier to classify real-world Chest X-rays and record their accuracy and AUROC score.

To set a control group for comparison, we want to see how original Diffusion model performs on the task. We generated the same number of un-fine-tuned samples and put them into same experiment setting. Moreover, we also trained the same classification model with the same frontal Chest X-ray dataset and then test with the same set of synthetic frontal chest X-rays, and record accuracy and AUROC score performance of this classification model.

We ran all of the experiments 5 times and took the average performance. The results are presented as follows.

4.1.2 Results

Table 4.1 shows the result of our initial experiment setting. With this random sampled tuning set. We received pretty low performance in such classification task.

Number of Fine-tune Images	Accuracy	AUROC		
	Overall	Overall	Unhealthy	No Finding
No Fine tune	50%	0.397	0.421	0.371
1,000	54%	0.587	0.588	0.586

Table 4.1: Initial Experiment: Trained on synthetic data

Table 4.2 shows the result of our control group of our initial experiment. Though not perfectly performed, it presents a better performance for each of the metric we measured compared to the classifier trained on synthetic dataset.

Number of Fine-tune Images	Accuracy	AUROC		
	Overall	Overall	Unhealthy	No Finding
1,000	61%	0.677	0.686	0.669

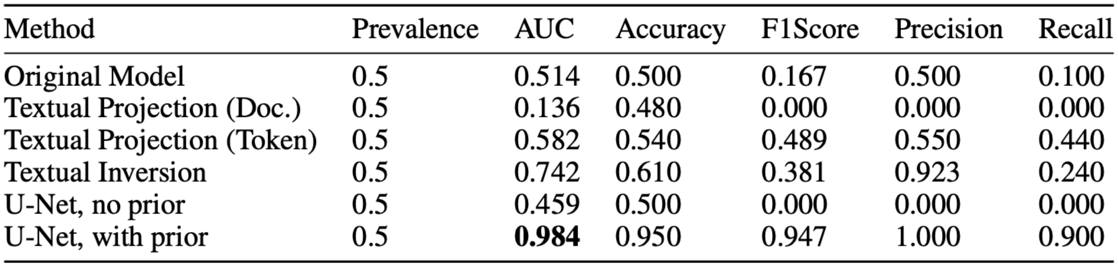
Table 4.2: Initial Experiment Control Group: Trained on real data

4.1.3 Discussion

From the result of our initial experiment, we see that fine-tuning as a domain-adaption technique to transfer learning weights from original model to targeted domain, as there presents a significant increase in AUROC performance of the two classifiers. AUROC, in particular, is one important metric to measure the accuracy and prediction performance of machine learning models [8].

However, we also observed a gap of performance metrics between the classifier trained on synthetic data and that trained on real-world data. This usually means that generative model does not fully capture the distribution of target domain. Thus, such domain-adaption technique can be insufficient or inefficient in generation tasks.

Moreover, recent works have shown promising results in generating medical-imaging-domain-specific synthetic images. For example, Chambon et al. presented a high performance statistics tasking frontal chest X-ray image classifier to classify and identify whether the synthetic Chest X-ray generated from fine-tuned latent diffusion model correspond to the label the generative model is asked to generate. Table 4.3 shows the results from their research. They concluded that a fine-tuning technique modeled with fine-tuning U-Net with prior are able to research the best performance [11], which is the same fundamental architecture compared to Dreambooth fine-tuning model [59]. Such results may amplify the gap between the classifier trained on synthetic dataset and the classifier trained on real data. We figured more researches in such gap and ways to minimize such gap need to be done, and ways to measure the gap are also required. Therefore, we proposed a new evaluation



Method	Prevalence	AUC	Accuracy	F1Score	Precision	Recall
Original Model	0.5	0.514	0.500	0.167	0.500	0.100
Textual Projection (Doc.)	0.5	0.136	0.480	0.000	0.000	0.000
Textual Projection (Token)	0.5	0.582	0.540	0.489	0.550	0.440
Textual Inversion	0.5	0.742	0.610	0.381	0.923	0.240
U-Net, no prior	0.5	0.459	0.500	0.000	0.000	0.000
U-Net, with prior	0.5	**0.984**	0.950	0.947	1.000	0.900

Table 4.3: Previous research's performance in classifying task using synthetic frontal chest X-ray as test set [11]

pipeline along with a new metric to better numerically present the gap, and continued our research on this topic by attempting to find efficient ways to minimize such gap. To simply refer to such gap in performance, we refer such gap as the Trainability Gap (T-Gap).

4.2 Measurement 1: Evaluation Pipeline

The observation from previous experiment reveals the existence of Trainability Gap(T-Gap). Good testing results on models pre-trained on real-world dataset tasked to classify synthetic dataset does not necessary secure good performance when training a classifier using synthetic dataset. However, as discussed in Chapter 2, one important use of synthetic dataset is its promising future use as a cheap and easy-obtainable replacement of real-world dataset to training large-scale machine learning models. Such utility requires good performance of models trained on synthetic dataset. Minimizing the T-Gap for synthetic datasets is one challenge. Research tools to resent and measure the T-Gap is in need. However, based on the knowledge of the author, current research hardly touched on related topic.

Therefore, based on the simple need to presenting performance comparison between models trained on real-world data and models trained on synthetic data, we constructed a 2-way evaluation pipeline constructs simple as follows:

1. Train a classifier model using synthetic dataset, and then test the trained classifier using real-world dataset.

2. Train a classifier model using real-world dataset, and then test the trained classifier using synthetic dataset.

Researchers are able to observe and compare the resulting performance (e.g accuracy and AUROC) of the two classifier models to get insights on how much the T-Gap can be for the synthetic dataset for model training purposes on the targeted domain. Figure 1.1 presents a diagram of flow of our new evaluation pipeline.

4.3 Measurement 2: Fine-tune Efficiency Ratio

Now that we have methods to evaluate the performance of the two classifiers in the evaluation pipeline and are able to observe the T-Gap by comparing resulting performance of the two classifier. However, it still maintains unclear on numerical measurement of the T-Gap. Therefore, we proposed a Fine-tune Efficiency Ratio (FTER) as follows to be able to numerically compare and get insights about the T-Gap for synthetic datasets generated through different methods.

$$FTER = \frac{\text{AUROC of classification task trained on synthetic data test of real-world data}}{\text{AUROC of classification task trained on real-world data test on synthetic data}}$$

FTER compares the AUROC of the two classifier trained in the evaluation pipeline and computes how similar the AUROC metric of the two models are. Recall that AUROC is a good indication of a model's performance in prediction [8]. FTER tells how similar the two models performs, and therefore tells how similar the model performance can be when trained on a synthetic dataset verses trained on real-world dataset. Detailed general analysis of how such FTER metrics work is shown in section 3.2.2.

4.4 Experiment 2: Naive Baseline

While proposing new measuring pipelines along with new metrics in examining the task of domain-specific synthetic dataset generation, this thesis is curious on how different methods perform under such set of measurements. To start with the experiments, we begin with a baseline as follows.

Recall that the goal of this thesis is to find efficient ways for generating domain-specific synthetic dataset, and, based on our literature review (chapter 2), model-fining is shown to be an efficient way of transfer learning for pre-trained models for domain-specific generation tasks. Then, we constructed a broad set up for our experiments. We fine-tune generation models with some tuning set using some fine-tuning model. Next, with the fine-tuned model, we generate 10,000 images with generated labels to train some classifier. Finally, we run through our proposed evaluation pipeline and compare the performance of the two classifier using the new metric, FTER, we proposed.

Such experiment set up greatly depends on the quality of some models (i.e. dataset, generation model, fine-tuning model and classifier model) to get good and conclusive results of the performance of each method. Based on our literature review, we picked a set of robust dataset and models as follows for our experiments.

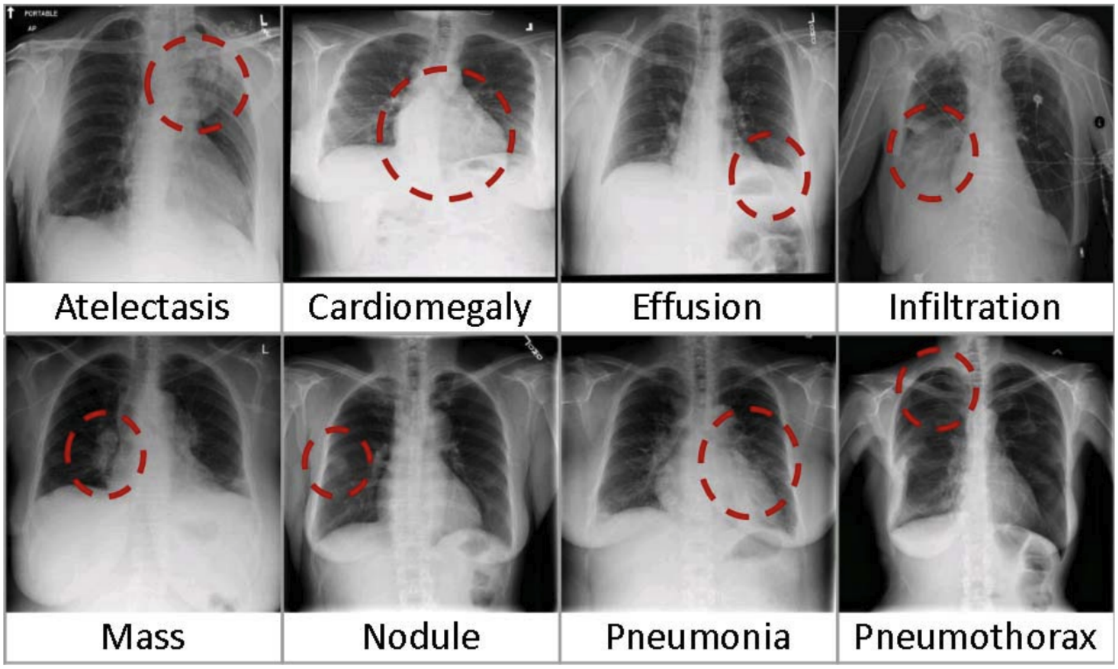


Figure 4.1: ChestX-ray14 dataset. Reproduced from [70].

NIH ChestX-ray14 dataset The dataset we mainly focused (choose our data from) on is the NIH ChestX-ray14 dataset [70] (Figure 4.1), which contains 112,120 frontal-view X-ray images of 30,805 patients. It is annotated by experts in frontal Chest X-ray domain with 14 different thoracic pathology labels and one no-finding label, along with some of the patient information, such as ages and genders. To simplify the generation task, we reclassifiy the X-rays dataset into a binary split with one label indicates some symptoms shown in the frontal chest X-rays and one label indicates no finding.

With fine-tuning generative model on a subset of the ChestX-ray14 dataset, we are able to reproduce relatively realistic results of frontal Chest X-rays. Figure 3.1 shows an example of comparison of synthetic Chest X-rays before and after we deploy fine-tuning technique on the stable diffusion image generation model.

Generative Model: Latent Diffusion As this thesis focused on the domain of frontal chest X-ray medical image generation, the task requires a robust pre-trained image generation model. Based on literature review (chapter 2), latent diffusion models show impressing ability ability in generating high-quality and high-resolution synthetic images [58], and researches also shown effective in fine-tuning diffusion models for tasking frontal chest X-ray generation [11]. We utilized the latent diffusion model as our main generative model in our experiments.

Though some research reports that prompting-based approaches greatly effect models' performances [31], from our experiment, we see that impact of good prompt does not show a strong impact on quality of synthetic dataset. Therefore, in our experiment settings, we stick with the single prompt formed as "an image of [label] chest X-ray" as the the tuning text guidance and generation text guidance for our tuning and generation process.

Fine-tune Model: Dreambooth As discussed in Chapter 2, the task of out-of-domain generation requires domain adaption technique, which model fine-tuning is one of the popular and efficient technique for such task. Among various models for fine-tuning, we found Dreambooth model is shown to be efficient in our desired task of domain adaption of frontal chest X-ray domain [11]. Thus, for our baseline, we integrate the Dreambooth model as our domain adaption technique for generating out-of-domain image datasets.

Classification Model: CheXNet As discussed briefly in section 3.2.1, for the task of classifying frontal chest X-rays, the model CheXNet [55] is shown to be efficient with great performance. Therefore, we implemented CheXNet as the classifier model in our evaluation pipeline in our experiments. We modified the model into the classification task of binary "Unhealthy" and "No Finding" cases for simpleness.

4.4.1 Experiment Setting

To begin our experiment with a baseline for comparison purposes, we started with a naive random sampling approach. We models a naive baseline task as follows:

1. We **randomly** picked n number of domain-specific frontal chest X-ray images from the recollected binary version (label of "Unhealthy" and "No Finding") of NIH ChestX-ray14 dataset.

2. We tune the latent diffusion image generation model with this set of data using Dreambooth model.

3. We then generate 10,000 images for each class, "Unhealthy" and "No Finding" (use word "Healthy" for prompting)), with the prompt of "a image of [label] chest X-ray".

4. Train a CheXNet model using the synthetic dataset. Test it using randomly picked 10,000 images from NIH ChestX-ray14 dataset and record its performance.

5. Train a CheXNet model using the same number of images for each class. Test it using randomly picked 10,000 images in total from the synthetic dataset and record its performance.

6. Compare and analysis the resulting performance of the two classifier.

Following the steps, we got the results as follows.

4.4.2 Results

We ran our baseline experiment based on steps above for 5 runs, and got the average performance result as follows.

Number of Fine-tune images	Accuracy			AUROC		
	Overall	Unhealthy	No Finding	Overall	Unhealthy	No Finding
No Fine tune	50%	100%	0%	0.397	0.421	0.371
500	52.5%	90%	25%	0.541	0.5429	0.540
1,000	54%	34%	73%	0.587	0.588	0.586
10,000	51%	72%	31%	0.478	0.479	0.478

Table 4.4: Random sampled tuning set: result trained on synthetic Chest X-rays

Number of fine tune images	Accuracy	AUROC	FTER
500	54%	0.575	0.939
1000	61%	0.677	0.866

Table 4.5: Random sampled tuning set: result trained on real Chest X-rays

Table 4.4 shows the resulting performance of the CheXNet classifier trained on synthetic dataset. Table 4.5 shows the resulting performance of the classifier trained on real-world data but test on the synthetic dataset.

4.4.3 Discussion

The results show a bell shaped curve with the best accuracy and AUROC score at around 1,000 random images for fine tuning. The results provide information that the size of the fine tuning dataset matters for area specific fine tuning for latent diffusion models and it is not naively "the more is better".

Such baseline experiment also confirmed our assumption with the existence of the T-Gap between the train and test performance of the synthetic dataset. We see from Table 4.4 and Table 4.5 that, when the tuning set is at 1,000, it reaches the highest average accuracy and AUROC metric in the classification task. However, it also has the higher deviation on FTER metric comparing to the perfect FTER of 1. This might be due to the reason that, when taken in the tuning set of other sizes, the model performs worse off and more like random classifier. The image quality of the synthetic dataset is not high enough for a pretrained-on-real classifier to identify the feature of its label class. As the quality of synthetic dataset gets higher, the T-Gap might be more obvious. We need more experiment and analysis for methods that can generate higher quality images.

4.5 Experiment 3: Spectral Clustering on Pixel

To touch on the problem of generating higher quality images, based on our literature review (Chapter 2), we take two approaches from better tuning data and more efficient model architecture to target the problem. We started with approaches on finding better tuning set. Specifically, we first focused on applying clustering techniques to find more efficient subset for task of domain adaption.

In terms of choosing the clustering algorithm, research show that spectral clustering algorithm shows to improve the data efficiency of machine learning models, and therefore resulting better model performance [36]. Thus, our experiment focused on the clustering algorithm of spectral clustering.

4.5.1 Experiment Setting

We applied the experiment steps as described in Experiment 2 (section 4.4). The only difference is that, instead of randomly sampling the tuning set from the binary ChestX-ray14 dataset, we applied clustering techniques. We first picked 10,000 random images from the full binary ChestX-ray14 dataset. We then cluster the 10,000 images based their converted RGB pixel representations using spectral clustering algorithm into 10 clusters for each label. We then pick equal number of images from each cluster to form our final tuning set. If the number of images of some cluster is not sufficient, we randomly duplicate the images in that cluster in order to enforce equal weights for each cluster, which eliminate the bias of data in some underrepresented clusters. Next, we put our picked tuning set into model fine-tuning, and run the process of the evaluation pipeline. We got the experiment results as follows.

4.5.2 Result

We run our experiment for 5 runs, and here is the average performance we get.

Number of Fine-tune images for each label	Accuracy	AUROC		
	Overall	Overall	Unhealthy	No Finding
500	52.66%	0.572	0.572	0.570
1,000	50.33%	0.546	0.549	0.543

Table 4.6: Spectral Clustering on pixel: Classifier trained on synthetic data

Number of Fine-tune images	Accuracy	AUROC			FTER
	Overall	Overall	Unhealthy	No Finding	Overall
500	57.67%	0.706	0.731	0.680	0.810
1,000	55.33%	0.676	0.690	0.662	0.808

Table 4.7: Spectral Clustering on pixel: Classifier trained on real, test on synthetic

4.5.3 Discussion

From our result, we see that the synthetic dataset deploying cluster-based tuning set selection based on RGB pixel representation of images in the fine-tuning process shows a very subtle improvement. The accuracy does not change much, while the AUROC is improving a bit, which might showing the model making slightly better predictions. However, the FTER metric drops, indicating a possibility of a broadened T-Gap while generating higher quality synthetic datasets.

We also see that, for different size of tuning set put into the fine-tuning model, FTER does not change much, indicating that the fine-tuning efficiency ratio (FTER) is possibly a model/method-wise metric that can used to examine how a generative model can generate synthetic dataset that performs similarly in either taking into training or taking as a testing set. More experiments on different methods and models are required. Desired model/method should have FTER near 1 along with high accuracy and overall AUROC.

4.6 Experiment 4: Spectral Clustering on CLIP

Research shows that the performance of spectral clustering can be improved when clustering on some deep embedding space [17]. From there, we shaped our next experiment.

Image embedding is one helpful technique for better performance in image clustering tasks. Using image embeddings in machine learning provides a range of benefits over pixel-based approaches. These embeddings distill semantic information from images into vectors, enabling models to understand content and context more effectively [17]. There are various different embedding spaces align with the image space. One of the most famous and shown greatly efficient and effective one is the CLIP embedding [54].

By learning joint representations of images and text, CLIP embedding space facilitates cross-modal learning, improving generalization and robustness across tasks. Furthermore, pre-trained embeddings support transfer learning, enhancing performance with limited labeled data. They reduced dimensionality, making models easier to learn (model usually learns better in lower dimension tasks) and also more computationally efficient. Embeddings are also more robust to variations in model decisions. Overall, image embeddings empower machine learning models with semantic understanding, cross-modal capabilities, and enhanced efficiency, advancing their applicability across diverse domains [54].

For the reasons stated above, we applied CLIP embedding to the images, and perform same method as Experiment 3 (secton 4.5).

4.6.1 Experiment Setting

We generally performed the same experiment technique as described in Experiment 3(section 4.5). Instead of clustering the images based on their RBG pixel representations, we first computed the CLIP embedding representation of the chosen images, and then perform the clustering and picking steps to form the tuning set.

4.6.2 Result

We run our experiment for 5 runs, and here is the average performance we get.

Number of Fine-tune images for each label	Accuracy	AUROC		
	Overall	Overall	Unhealthy	No Finding
500	54%	0.575	0.565	0.584
1,000	51%	0.541	0.548	0.535

Table 4.8: Spectral Clustering on CLIP: Classifier trained on synthetic data

Number of Fine Tune images	Accuracy	AUROC			FTER
	Overall	Overall	Unhealthy	No Finding	Overall
500	83%	0.935	0.940	0.930	0.615
1,000	79%	0.891	0.905	0.877	0.607

Table 4.9: Spectral Clustering on CLIP: Classifier trained om real, test on synthetic

We observe a slightly higher accuracy and AUROC for the classifier trained on synthetic dataset, and a significant increase in the accuracy and AUROC metric for the classifier trained on real dataset but test on synthetic data. The FTERs deviate more from 1, but are still similar for different size of tuning set.

4.6.3 Discussion

The result shows a significant increase in both the accuracy (from ~55% to ~80%)and AUROC (from ~0.68 to ~0.91) of the classifier model trained on real dataset and test on synthetic dataset. This means that the classifier model is able to capture and classify synthetic images based on the characteristic it generated. Therefore, the fine-tuned model is generating better quality images.

The classifier trained on the synthetic images, on the other hand, also shows improvement in the metrics, but the growth is a lot less compared to the classifier trained on real dataset. This is also shown from the decrease in FTER, meaning that the T-Gap on the synthetic dataset become bigger as the quality of synthetic images go up. We proposed that this might be because, when generating domain-specific images, the fine-tuned model does not generate datasets with various features that contains enough variations to train another model to identify the underlying data distribution. The generated dataset might stick too much to some tuning images. If this is the case, we may see better results when we modified the fine-tuning process to let the model learn more on the variations of the dataset. This idea shaped our experiment in section 4.8 on active data selection to enforce model learning more on variations of the tuning set.

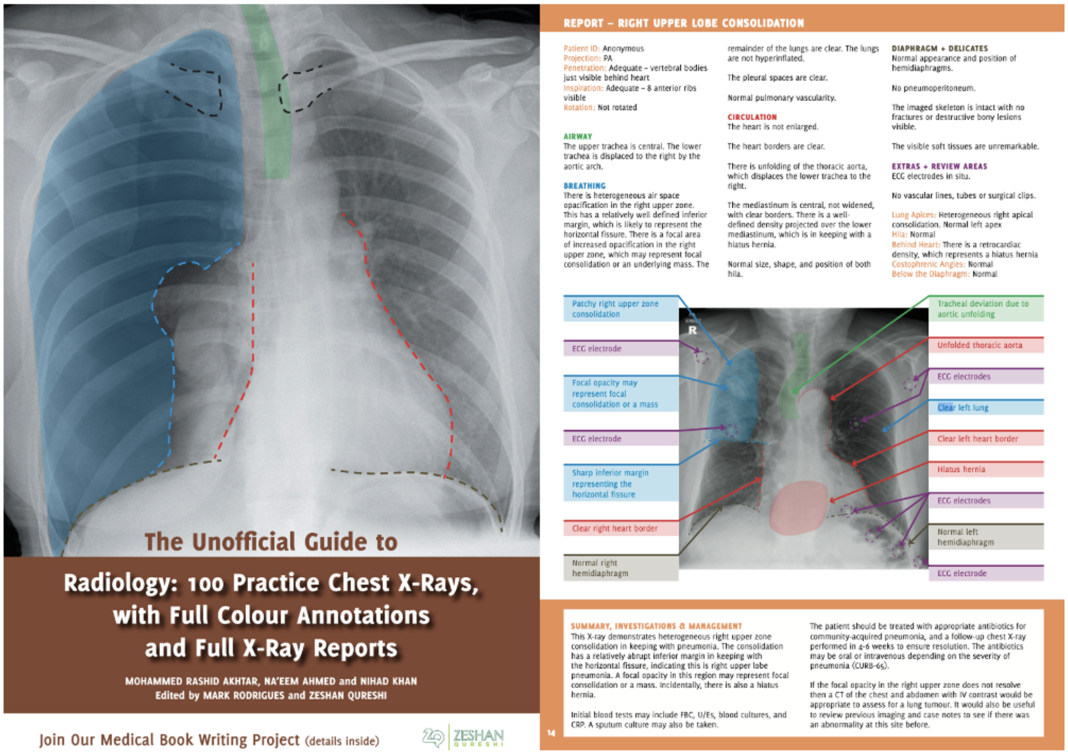
For the result of this experiment, we see that FTER still does not change much with different size of tuning set. We are more confirmed that FTER is a model/method-wise metric that can be calculated to numerically value a model/method's performance in synthetic dataset generation for re-training purposes.

4.7 Experiment 5: Textbook Dataset

As mentioned in section 3.2.1, we aimed to find better quality data from source. To target on the question of what data can have higher quality and be more informative to learn, we made an assumption that more educational data in textbooks might be more informative and thus possible can obtain better training performance. Therefore, we tested for an experiment specially trained on our originally collected textbook dataset.

4.7.1 Experiment Setting

We manually collected over 1,000 frontal Chest X-ray image samples from 5 high-rated textbooks on chest X-ray radiology ("The Chest X-Ray: A Survival Guide" [15], "The Unofficial Guide to Radiology: 100 Practice Chest X-rays, with Full Colour Annotations, and Full X-Ray Reports" [25], "Felson's Principles of Chest Roentgenology: A Programmed Text"



The Unofficial Guide to

Radiology: 100 Practice Chest X-Rays, with Full Colour Annotations and Full X-Ray Reports

MOHAMMED RASHID AKHTAR, NA'EEM AHMED and NIHAD KHAN
Edited by MARK RODRIGUES and ZESHAN QURESHI

Join Our Medical Book Writing Project (details inside)

ZESHAN QURESHI

REPORT – RIGHT UPPER LOBE CONSOLIDATION

Patient ID: Anonymous
Projection: PA
Penetration: Adequate – vertebral bodies just visible behind heart
Inspiration: Adequate – 8 anterior ribs visible
Rotation: Not rotated

AIRWAY
The upper trachea is central. The lower trachea is displaced to the right by the aortic arch.

BREATHING
There is heterogeneous air space opacification in the right upper zone. This has a relatively well defined inferior margin, which is likely to represent the horizontal fissure. There is a focal area of increased opacification in the right upper zone, which may represent focal consolidation or an underlying mass. The remainder of the lungs are clear. The lungs are not hyperinflated.

The pleural spaces are clear.

Normal pulmonary vascularity.

CIRCULATION
The heart is not enlarged.

The heart borders are clear.

There is unfolding of the thoracic aorta, which displaces the lower trachea to the right.

The mediastinum is central, not widened, with clear borders. There is a well-defined density projected over the lower mediastinum, which is in keeping with a hiatus hernia.

Normal size, shape, and position of both hila.

DIAPHRAGM + DELICATES
Normal appearance and position of hemidiaphragms.

No pneumoperitoneum.

The imaged skeleton is intact with no fractures or destructive bony lesions visible.

The visible soft tissues are unremarkable.

EXTRAS + REVIEW AREAS
ECG electrodes in situ.

No vascular lines, tubes or surgical clips.

Lung Apices: Heterogeneous right apical consolidation. Normal left apex
Hila: Normal
Behind Heart: There is a retrocardiac density, which represents a hiatus hernia
Costophrenic Angles: Normal
Below the Diaphragm: Normal

Patchy right upper zone consolidation
ECG electrode
Focal opacity may represent focal consolidation or a mass
ECG electrode
Sharp inferior margin representing the horizontal fissure
Clear right heart border
Normal right hemidiaphragm
R
Tracheal deviation due to aortic unfolding
Unfolded thoracic aorta
ECG electrodes
Clear left lung
Clear left heart border
Hiatus hernia
ECG electrodes
Normal left hemidiaphragm
ECG electrode

SUMMARY, INVESTIGATIONS & MANAGEMENT
This X-ray demonstrates heterogeneous right upper zone consolidation in keeping with pneumonia. The consolidation has a relatively abrupt inferior margin in keeping with the horizontal fissure, indicating this is right upper lobe pneumonia. A focal opacity in this region may represent focal consolidation or a mass. Incidentally, there is also a hiatus hernia.

Initial blood tests may include FBC, U/Es, blood cultures, and CRP. A sputum culture may also be taken.

The patient should be treated with appropriate antibiotics for community-acquired pneumonia, and a follow-up chest X-ray performed in 4-6 weeks to ensure resolution. The antibiotics may be oral or intravenous depending on the severity of pneumonia (CURB-65).

If the focal opacity in the right upper zone does not resolve then a CT of the chest and abdomen with IV contrast would be appropriate to assess for a lung tumour. It would also be useful to review previous imaging and case notes to see if there was an abnormality at this site before.

14

Figure 4.2: Example of Textbook we collected data from. Reproduced from [25].

[21], "Chest X-Ray Made Easy" [14], "Muller's Imaging of the Chest: Expert Radiology Series" [66]), and made up our own textbook dataset. Figure 4.2 shows an example textbook we collected our data from. We collected image data as that on the center of the right page into our dataset.

Limited to the author's knowledge on medical text and chest X-rays, during our data collection process, we do not manually differentiate between the labels of each chest X-ray from the textbooks. Instead, we trained a ChestXNet classifier model [55] with 90% of the full NIH ChestX-ray14 dataset, which is around 100,000 image samples in total. We tested the ChestXNet classifier's performance on differentiating between real chest X-rays and found a 93% accuracy in such task, showing effectiveness in deploying the classifier model to label our textbook data. Thus, after we collectively capture all frontal chest X-rays from the data, we utilized this pre-trained ChestXNet classifier to label the textbook images. After labeling, we ended up with around 500 images for each of the label ("Unhealthy" and "No Finding"). We then put all off the selected data into tuning the latent diffusion image generation model and run through our evaluation pipeline.

4.7.2 Result

We run our experiment for 5 runs, and here is the average performance we get.

Number of Fine-tune images for each label	Accuracy	AUROC		
	Overall	Overall	Unhealthy	No Finding
~500	50%	0.567	0.567	0.567

Table 4.10: Textbook Dataset: Classifier trained on synthetic data

Number of Fine-tune images	Accuracy	AUROC			FTER
	Overall	Overall	Unhealthy	No Finding	Overall
~500	81%	0.893	0.899	0.887	0.635

Table 4.11: Textbook Dataset: Classifier trained on real, test on synthetic

We observe the slight increase in AUROC but also a slight drop in accuracy in the classifier train on synthetic dataset compared to our baseline. For the model trained on the real dataset but test on the synthetic images, we see a great increase in both accuracy and AUROC score. The performance is reaching similar results as the result we get from applying spectral clustering methods on CLIP embedding for selecting the tuning set (Experiment 4.4). We know that the model is indeed generating images with better quality.

4.7.3 Discussion

Limited to the number of chest X-ray images we are able to collect from medical textbooks, we only performs this experiment setting with the fine-tuning set of around 500 number of images for each of the labels. But we can still get insights on how good quality data matters in the task of synthetic dataset generation for re-training purposes. As we see from the result of the classifier trained on real images, the accuracy and AUROC is pretty robust there.

However, we met a same problem as that presented in Experiment 4.4. The performance of the classifier trained on synthetic images does not improve that much and even the accuracy of the classifier drops. Though FTER get closer to 1, it is simply because the classifier trained on real data performs a bit worse off in such experiment setting, which is not the case we desired. We suppose that it might be because images chosen in textbook are too much "typical" cases of frontal chest X-rays, which lack of certain variations cross the dataset to let the fine-tuning model fully learn the distribution of targeted domain. This confirms our thoughts on learning the maximum variations of the tuning set, and forms our next experiment on active data selection methods to enforce the fine-tuning model selectively learn tuning data with most variations.

4.8 Experiment 6: Active Data Selection

As discussed in section 2.2.2 and section 3.3.2, active data selection is one method that can increase model efficiently by tasking a third-party oracle to label and select data taken into training. With the experiment observations above, we found the possibility of model not fully able to learn and reproduce data variations. Thus, we modeled an experiment setting embedding a active-learning oracle enforcing model to learn data with more variation into the Dreambooth fine-tuning model. Detailed experiment settings as follows.

4.8.1 Experiment Setting

Consider the general idea of active learning. During the training process, one oracle decides whether a data will be taken into training. In our case, we want to use activate data selection technique to enforce the model learning more cross the variation about the data, which means we need to enforce the model to learn samples that it is currently unconfident to generate. Ideally, we desire an oracle to query the current model state whether it know about the current training batch of data. However, directly querying a model can be hard. But what can be easily done to construct an oracle to compare the similarity of the generated results of data of input labels of current training batch. The similarity can show how much of the image might be able to be reproduce and thus might have been learned. Previous research also shown that similarity-based learning is effective in improving model performance [51, 50]. When we know how similar the current knowledge and the training batch is, we may set up a similarity threshold to decide whether to take the training batch will be taken into training. Data above the threshold is categorized as too similar with less variance that has higher probability that is learned, while data samples below the threshold is categorized as unlearned data with more variance.

Therefore, the general idea of the algorithm is, for each batch, generate some samples of the knowledge of model's current state and then compute a similarity score with the current training batch of data. If the similarity score is above a certain threshold, then we neglect this training batch. In this way, the model can learn data batches in the dataset that are unsimilar and thus with more vairance compared to its knowledge, and self-reweight the training data to make the learning process more efficient.

Algorithm 1 shows a pseudocode of the idea of active data selection embedded Dreambooth model. In our experiment, we pre-applied the clustering on CLIP technique in Experiment 4 (section 4.4) for the tuning set, and we used cosine similarity score as our similarity measure and threshold of 0.3, while other options for the set of measures (different metric for similarity score and different thresholds) is available.

Algorithm 1 Active data selection embedded Dreambooth model idea

```
**Data:** Pre-processed tuning set using cluster on CLIP technique in Experiment 4
for each epoch do
    for each batch of $n$ data do                          ▷ $n$ is the size of each batch
        CurGen ← CLIP representations of $n$ images generated by current model
        DataBatch ← CLIP representations of the current batch
        SimilarityScore ← the cosine similarity of CurGen and DataBatch
        if SimilarityScore $> 0.3$ then
            The model continues without learning this batch.
        else if otherwise then
            Performs Dreambooth fine-tuning
        end if
    end for
end for
```

We implemented such algorithmical idea on Dreambooth model, and the following shows the results of our experiment.

4.8.2 Result

We run our experiment for 5 runs, and here is the average performance we get.

Number of Fine-tune images for each label	Accuracy	AUROC		
	Overall	Overall	Unhealthy	No Finding
500	60.33%	0.637	0.646	0.628
1,000	56%	0.591	0.595	0.587

Table 4.12: Active Data Selection: Classifier trained on synthetic data

# of fine tune images	Accuracy	AUROC			FTER
	Overall	Overall	Unhealthy	No Finding	Overall
500	89%	0.957	0.961	0.943	0.666
1,000	83.66%	0.923	0.930	0.916	0.640

Table 4.13: Active Data Selection: Classifier trained on real, test on synthetic

We observe increase in both accuracy and AUROC in this experiment. FTER is have also increased compared to other methods, indicating such experiment model shows fair performance in the task of domain-specific synthetic data generation for model training purposes.

Figure 4.3 and Figure 4.4 shows some examples of our synthetic frontal chest X-rays deploying this active data selection approach. We see pretty decent results of high-quality synthetic frontal chest X-rays.

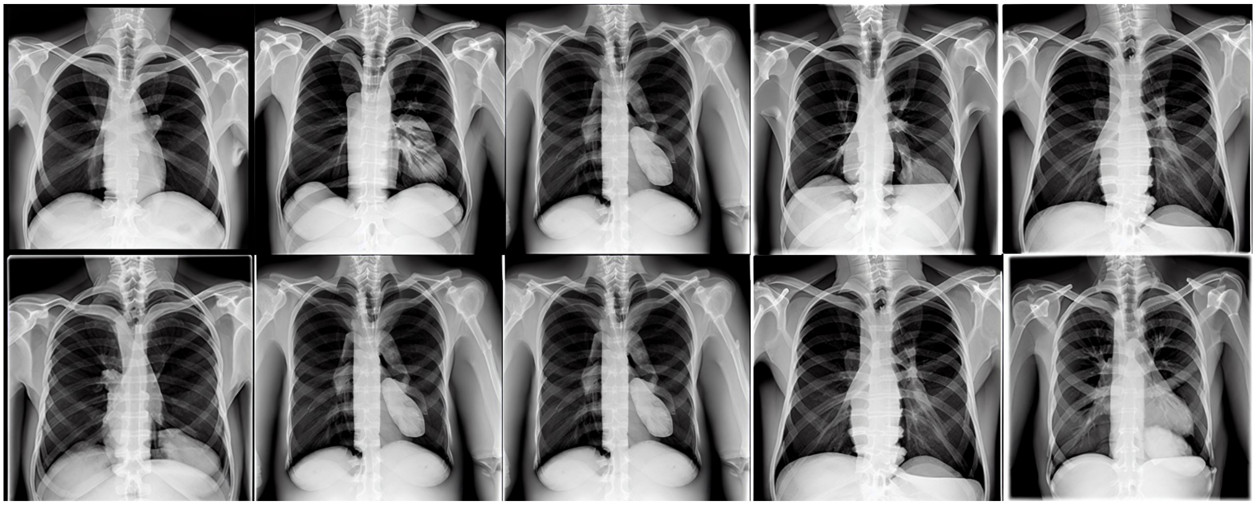


Figure 4.3: Example of synthetic frontal Chest X-rays with active data selection method: label "Unhealthy"

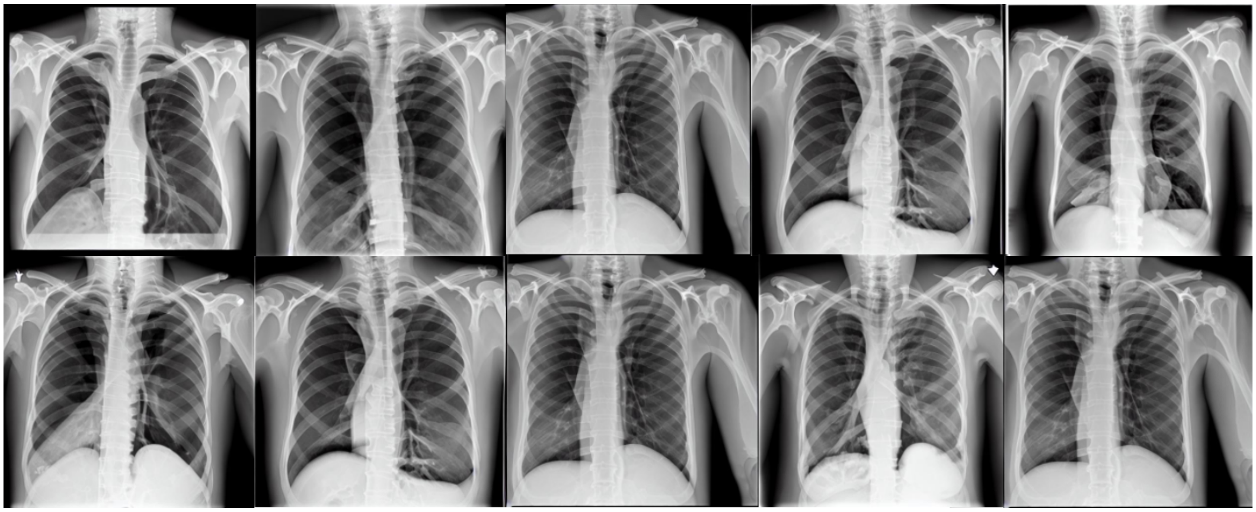


Figure 4.4: Example of synthetic frontal Chest X-rays with active data selection method: label "No Finding"

4.8.3 Discussion

From the experiment results, we see an increase in accuracy, AUROC and FTER metrics, indicating that the model is more efficient and researching better performance under this experiment set up. Thus, an active-data-selection based approach for domain adaption along with a cluster-based pre-processing seems to be more efficient under our evaluation pipeline and metrics for the task of domain-specific synthetic dataset generation for model training purposes.

Such improvement in performance somehow confirms our idea that, on the task for fine-tuning model for domain adaption for synthetic dataset generation for model training purposes, better performance requires variance-emphasized training, which explicitly enable the model to capture the variance across the underlying distribution of the target domain. The increase in FTER shows that such approach can decrease the T-Gap of synthetic dataset.

In experiment, we also observe that, regardless of the tuning data size, FTER computes fairly similar results depend on models. Recall the results from Experiment 2, 3, 4 (section 4.4, 4.5, 4.8), for each of the experiment model set up, FTER gives similar results. Now we are able to conclude that FTER is a model/method-wise metric that can be used to identify and measure the model-wise T-Gap for the synthetic dataset generated by a specific generative model on a specific domain.

4.9 Insights

After meticulously conducting and analyzing the outcomes of our six comprehensive experiments, we confidently assert that employing an active-data-selection based approach coupled with a cluster-based pre-processing technique constitutes a superior method for fine-tuning models as a domain adaptation strategy. Specifically, this method proves highly effective in facilitating the adaptation of image generation models to specific out-of-distribution domains, thereby enhancing synthetic dataset generation for the purpose of model training.

In addition to the aforementioned findings, our investigations reveal a distribution shift when transitioning from a tuning set size of 1,000 to 500, particularly when comparing against the naive baseline of random sampling. This shift underscores the tangible impact of implementing various strategies aimed at improving model efficiency.

Moreover, the results shows effectiveness of our proposed new metric, termed Fine-tune Efficiency Ratio (FTER), which presents a model-specific measure capable of efficiently gauging the Trainability Gap (T-Gap) inherent within synthetic datasets targeted for generative models. The introduction of FTER represents a significant advancement, providing researchers with a valuable tool for precisely assessing and addressing the challenges associated with the T-Gap phenomenon. We anticipate widespread adoption of our evaluation pipeline and the utilization of FTER across diverse research endeavors, as they promise to furnish invaluable insights and viable solutions for mitigating the T-Gap problem, thereby advancing the efficacy and applicability of generative models in various domains.

Chapter 5

Conclusions

5.1 Recap

Recap from the background and goal of this thesis. As, in the realm of machine learning, model training requires more and more high-quality dataset to achieve better performances. There reveals the problem of data paucity. As generative models shows their ability in generating high-quality data pieces, synthetic datasets might be able to serve as a possible solution for the data paucity problem. But there still remains a problem that current generative models has domain restrictions based on the domain of the dataset they were trained on. Out-of-domain synthetic dataset generation still remains a challenge. There are techniques as model fine-tuning that can efficiently transfer learning of generative models to adapt the targeted domain, and have shown to produce synthetic data with high fidelity. However, there is not much research focusing on how such out-of-domain synthetic dataset produced with adapted generative model can perform when put into training other machine learning models, which is the question that we care about to deploy synthetic dataset as a solution to the data paucity problem in real world scenarios. This thesis targeted at this problem. This thesis focuses on the research question on how domain-specific synthetic dataset can be used for model training purposes.

To investigate on the problem, we proposed a 2-way evaluation pipeline, which involves training and testing on two domain-specific classifier model (One is trained on synthetic dataset and test on real-world dataset, while the other is trained on real-world dataset and test on synthetic dataset.) and comparing their performance. Such comparison can tell how the synthetic dataset performs when deployed as train set and as test set. From our experiments, we observed the existence of a difference/gap between the two pipes. Classifiers trained on synthetic dataset generally performs worse off, indicating the synthetic dataset produced by adapted model does not fully reproduce the underlying distribution of the targeted domain. There is a gap between how well synthetic datasets can be trained and how well synthetic datasets can reproduce the features of targeted domain. We termed such gap as Trainability Gap (T-GAP).

To numerically evaluate the T-Gap, we proposed the Fine-tune Efficiency Radio (FTER) metric, which is the ratio of AUROC metric of the two domain-specific classifier in the 2-way evaluation pipeline. FTER can tell how similar the performances of models are when taken the synthetic dataset as train set and test set. The closer FTER is to 1, the higher the similarity is in their performance. FTER is a model/method-wise metric, which means that it can tell the performance of adapted generative model in the task of domain-specific synthetic dataset generation. Good models should have high accuracy in the domain-specific classifier along with FTER close to 1.

We deploy the 2-way pipeline and FTER metric to examine on different experiment setting to figure out which methods can be more efficient in increasing performance and efficiency of domain-adapted generative models. We targeted on the approach of better data, which focused on cluster-based preprocessing of training data and collecting better data from source, and better architecture, which focused on active-learning-based methods. From our experiment results, we find out that an approach of embedding a active-data selection architecture into the domain-adaption model along with a cluster-based data preprocessing produces shows to be most efficient with best performance and less training data under the 2-way pipeline and FTER metric.

5.2 Limitations

In this thesis, we present limitations pertaining to both the dataset employed and the model utilized. These constraints warrant careful consideration to accurately interpret and contextualize our findings.

First, regarding the domain of the dataset, one prominent limitation lies in its size and representativeness. This thesis focused on the domain of medical imaging, in particular about frontal chest X-rays. Though we observed a good outcome of our method, it is still unclear if the results is also application to other domains. This limitation might constrain the applicability of our findings in diverse domains. More investigation cross different domains is needed.

Second, the model we deployed present certain limitations. This thesis specifically focused on methods around the image generation model latent diffusion and the fine-tuning model Dreambooth. As discussed in Chapter 2, there exists various options for generative models and domain adaption techniques. Though our experiment results shows existence of T-Gap and present efficient methods in minimizing such gap, whether such solution is a model-specific or transferable remains still unclear. More research and experiments cross different models is required.

Moreover, for the methods of touching the task of efficient fine-tuning, this thesis only touched on several selected approaches, which presents limitations as more efficient approaches out of the scope of this thesis may exist.

In conclusion, while our research endeavors have yielded valuable insights on how well synthetic dataset can be used for model training, it is imperative to acknowledge and address the inherent limitations associated with the dataset and model employed in our experiments. By transparently delineating these constraints, we can enhance the credibility and utility of our research outcomes while paving the way for future investigations to build upon our work and overcome these limitations.

5.3 Future Research

For future researches in the short run, it is evident that further research is imperative to advance our understanding and address the limitations inherent in our current study. This necessitates exploring additional datasets focusing on diverse domains, researching of such task deploying different generative models, and featuring such result with more approaches in efficient fine-tuning. The above three aspects serve as fundamental ways to touch on the limitations of this research thesis and produce a broader view of the phenomenon of existence of T-Gap.

Moving more forward, in addition to diversifying domain-specific datasets and models, future research should prioritize enhancing interpretability of the existance of T-Gap. By leveraging more domain-specific data into such task taking the advantage of the 2-way evaluation pipeline we proposed, researches are able to investigate more around T-Gap between synthetic dataset for training and testing, and enhance the validity and reliability of their analyses of the quality of the synthetic dataset. Then, with more results, researchers should be able to get more understanding on such complex phenomena, and then propose better possible reasons and solutions to resolve it.

Once the T-Gap is reasonably diminished, we foresee the efficient application of utilizing domain-adaption techniques to adapt power pre-trained generative model into specific domain (e.g. medicine, finance, etc.) and generate high-quality synthetic dataset that are able to train machine learning models with fair performance. Since the task of generating synthetic dataset is low-cost, such synthetic datasets can greatly expand our existing dataset and resolve the massive data requirement to train some powerful machine learning models, especially on domains that is currently lack of data. This means that, envisioned within this framework is the prospect of training and deploying machine learning models across multiple domains, heralding a transformative paradigm shift in enhancing human capabilities across a myriad of domain-specific tasks.

References

[1] Amina Adadi. 'A survey on data-efficient algorithms in big data era'. In: *Journal of Big Data* 8.1 (2021), p. 24.

[2] Md Abul Bashar and Richi Nayak. 'Active learning for effectively fine-tuning transfer learning to downstream task'. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 12.2 (2021), pp. 1–24.

[3] Shai Ben-David et al. 'Analysis of representations for domain adaptation'. In: *Advances in neural information processing systems* 19 (2006).

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[5] Zalán Bodó, Zsolt Minier and Lehel Csató. 'Active learning with clustering'. In: *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*. JMLR Workshop and Conference Proceedings. 2011, pp. 127–139.

[6] Rishi Bommasani et al. 'On the Opportunities and Risks of Foundation Models'. In: *CoRR* abs/2108.07258 (2021). arXiv: `2108.07258`. URL: `https://arxiv.org/abs/2108.07258`.

[7] Sam Bond-Taylor et al. 'Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (Nov. 2022), pp. 7327–7347. ISSN: 1939-3539. DOI: `10.1109/tpami.2021.3116668`. URL: `http://dx.doi.org/10.1109/TPAMI.2021.3116668`.

[8] Andrew P. Bradley. 'The use of the area under the ROC curve in the evaluation of machine learning algorithms'. In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159. ISSN: 0031-3203. DOI: `https://doi.org/10.1016/S0031-3203(96)00142-2`. URL: `https://www.sciencedirect.com/science/article/pii/S0031320396001422`.

[9] Gustav Bredell et al. *Explicitly Minimizing the Blur Error of Variational Autoencoders*. 2023. arXiv: `2304.05939 [cs.CV]`.

[10] Lukas Budach et al. *The Effects of Data Quality on Machine Learning Performance.* 2022. arXiv: 2207.14529 [cs.DB].

[11] Pierre Chambon et al. *Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains.* 2022. arXiv: 2210.04133 [cs.CV].

[12] Pierre Chambon et al. *RoentGen: Vision-Language Foundation Model for Chest X-ray Generation.* 2022. arXiv: 2211.12737 [cs.CV].

[13] Wei-Lin Chiang et al. 'Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks'. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 257–266. ISBN: 9781450362016. DOI: 10.1145/3292500.3330925. URL: https://doi.org/10.1145/3292500.3330925.

[14] Jonathan Corne and Maruti Kumaran. *Chest X-Ray Made Easy E-Book: Chest X-Ray Made Easy E-Book.* Elsevier Health Sciences, 2015.

[15] Gerald De Lacey, Simon Morley and Laurence Berman. *The chest X-ray: a survival guide.* Elsevier Health Sciences, 2012.

[16] Jesse Dodge et al. *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping.* 2020. arXiv: 2002.06305 [cs.CL].

[17] Liang Duan et al. 'Improving Spectral Clustering with Deep Embedding and Cluster Estimation'. In: *2019 IEEE International Conference on Data Mining (ICDM).* 2019, pp. 170–179. DOI: 10.1109/ICDM.2019.00027.

[18] Abolfazl Farahani et al. *A Brief Review of Domain Adaptation.* 2020. arXiv: 2010.03978 [cs.LG].

[19] George Forman and Bin Zhang. 'Distributed data clustering can be efficient and exact'. In: *ACM SIGKDD explorations newsletter* 2.2 (2000), pp. 34–38.

[20] Ian J. Goodfellow et al. *Generative Adversarial Networks.* 2014. arXiv: 1406.2661 [stat.ML].

[21] Lawrence R Goodman. *Felson's principles of chest roentgenology, a programmed text.* Elsevier Health Sciences, 2014.

[22] Venkat Gudivada, Amy Apon and Junhua Ding. 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations'. In: *International Journal on Advances in Software* 10.1 (2017), pp. 1–20.

[23] Suriya Gunasekar et al. *Textbooks Are All You Need.* 2023. arXiv: 2306.11644 [cs.CL].

[24] Seokhyeon Ha, Sunbeom Jung and Jungwoo Lee. *Domain-Aware Fine-Tuning: Enhancing Neural Network Adaptability*. 2024. arXiv: 2308.07728 [cs.LG].

[25] Ali BAK Al-Hadithi and Zeshan Qureshi. *The Unofficial Guide to Radiology: 100 Practice Chest X-rays*. Elsevier Health Sciences, 2023.

[26] Douglas M Hawkins. 'The problem of overfitting'. In: *Journal of chemical information and computer sciences* 44.1 (2004), pp. 1–12.

[27] Jeremy Howard and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. 2018. arXiv: 1801.06146 [cs.CL].

[28] Jiayuan Huang et al. 'Correcting Sample Selection Bias by Unlabeled Data'. In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt and T. Hoffman. Vol. 19. MIT Press, 2006. URL: https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf.

[29] Stanisław Jastrzebski, Damian Leśniak and Wojciech Marian Czarnecki. *How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks*. 2017. arXiv: 1702.02170 [cs.CL].

[30] Ahmadreza Jeddi, Mohammad Javad Shafiee and Alexander Wong. *A Simple Fine-tuning Is All You Need: Towards Robust Deep Learning Via Adversarial Fine-tuning*. 2020. arXiv: 2012.13628 [cs.CV].

[31] Woojeong Jin et al. *A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models*. 2022. arXiv: 2110.08484 [cs.CV].

[32] Eun Seo Jo and Timnit Gebru. 'Lessons from archives: strategies for collecting sociocultural data in machine learning'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. ACM, Jan. 2020. DOI: 10.1145/3351095.3372829. URL: http://dx.doi.org/10.1145/3351095.3372829.

[33] Jared Kaplan et al. 'Scaling Laws for Neural Language Models'. In: *CoRR* abs/2001.08361 (2020). arXiv: 2001.08361. URL: https://arxiv.org/abs/2001.08361.

[34] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[35] Ivan Kobyzev, Simon J.D. Prince and Marcus A. Brubaker. 'Normalizing Flows: An Introduction and Review of Current Methods'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (Nov. 2021), pp. 3964–3979. ISSN: 1939-3539. DOI: 10.1109/tpami.2020.2992934. URL: http://dx.doi.org/10.1109/TPAMI.2020.2992934.

[36] Pavel Kolev and Kurt Mehlhorn. *Approximate Spectral Clustering: Efficiency and Guarantees.* 2018. arXiv: `1509.09188 [cs.DM]`.

[37] Harlan Krumholz, Sharon Terry and Joanne Waldstreicher. 'Data Acquisition, Curation, and Use for a Continuously Learning Health System'. In: *JAMA* 316 (Sept. 2016). DOI: `10.1001/jama.2016.12537`.

[38] Ananya Kumar et al. *Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution.* 2022. arXiv: `2202.10054 [cs.LG]`.

[39] JoonHo Lee and Gyemin Lee. *Feature Alignment by Uncertainty and Self-Training for Source-Free Unsupervised Domain Adaptation.* 2023. arXiv: `2208.14888 [cs.CV]`.

[40] Xinyu Lin et al. *Data-efficient Fine-tuning for LLM-based Recommendation.* 2024. arXiv: `2401.17197 [cs.IR]`.

[41] Mingsheng Long et al. 'Transfer Feature Learning with Joint Distribution Adaptation'. In: *2013 IEEE International Conference on Computer Vision.* 2013, pp. 2200–2207. DOI: `10.1109/ICCV.2013.274`.

[42] Mingsheng Long et al. 'Unsupervised domain adaptation with residual transfer networks'. In: *Advances in neural information processing systems* 29 (2016).

[43] Shayne Longpre et al. *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, Toxicity.* 2023. arXiv: `2305.13169 [cs.CL]`.

[44] George D Magoulas and Andriana Prentza. 'Machine learning in medical applications'. In: *Advanced course on artificial intelligence.* Springer, 1999, pp. 300–307.

[45] Jose G Moreno-Torres et al. 'A unifying view on dataset shift in classification'. In: *Pattern recognition* 45.1 (2012), pp. 521–530.

[46] Andrew Ng and Michael Jordan. 'On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes'. In: *Advances in Neural Information Processing Systems.* Ed. by T. Dietterich, S. Becker and Z. Ghahramani. Vol. 14. MIT Press, 2001. URL: `https://proceedings.neurips.cc/paper_files/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf`.

[47] Hieu T Nguyen and Arnold Smeulders. 'Active learning using pre-clustering'. In: *Proceedings of the twenty-first international conference on Machine learning.* 2004, p. 79.

[48] George Papamakarios et al. 'Normalizing flows for probabilistic modeling and inference'. In: *J. Mach. Learn. Res.* 22.1 (Jan. 2021). ISSN: 1532-4435.

[49] Junyi Peng et al. 'An Attention-Based Backend Allowing Efficient Fine-Tuning of Transformer Models for Speaker Verification'. In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. 2023, pp. 555–562. DOI: 10.1109/SLT54892.2023.10022775.

[50] Meixin Peng, Zhanshan Li and Xin Juan. 'Similarity-based domain adaptation network'. In: *Neurocomput.* 493.C (July 2022), pp. 462–473. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.12.089. URL: https://doi.org/10.1016/j.neucom.2021.12.089.

[51] Pedro O Pinheiro. 'Unsupervised domain adaptation with similarity learning'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8004–8013.

[52] Joaquin Quionero-Candela et al. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN: 0262170051.

[53] Alec Radford et al. 'Improving language understanding by generative pre-training'. In: (2018).

[54] Alec Radford et al. 'Learning Transferable Visual Models From Natural Language Supervision'. In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[55] Pranav Rajpurkar et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 2017. arXiv: 1711.05225 [cs.CV].

[56] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV].

[57] Danilo Jimenez Rezende and Shakir Mohamed. *Variational Inference with Normalizing Flows*. 2016. arXiv: 1505.05770 [stat.ML].

[58] Robin Rombach et al. 'High-Resolution Image Synthesis with Latent Diffusion Models'. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. DOI: 10.1109/cvpr52688.2022.01042. URL: http://dx.doi.org/10.1109/cvpr52688.2022.01042.

[59] Nataniel Ruiz et al. 'DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation'. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 22500–22510. URL: https://api.semanticscholar.org/CorpusID:251800180.

[60] Christoph Schuhmann et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. 2022. arXiv: 2210.08402 [cs.CV].

[61] Burr Settles. 'Active learning literature survey'. In: (2009).

[62] Mohammad Shehab et al. 'Machine learning in medical applications: A review of state-of-the-art methods'. In: *Computers in Biology and Medicine* 145 (2022), p. 105458.

[63] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: `1503.03585 [cs.LG]`.

[64] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: `1706.03762 [cs.CL]`.

[65] Roy Voetman, Maya Aghaei and Klaas Dijkstra. *The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Detection Models*. 2023. arXiv: `2306.09762 [cs.CV]`.

[66] Christopher M Walker and Jonathan H Chung. *Muller's Imaging of the Chest: Expert Radiology Series*. Elsevier Health Sciences, 2018.

[67] Jindong Wang et al. 'Generalizing to Unseen Domains: A Survey on Domain Generalization'. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. IJCAI-2021. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021. DOI: `10.24963/ijcai.2021/628`. URL: `http://dx.doi.org/10.24963/ijcai.2021/628`.

[68] Jing Wang et al. 'Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by Gaussian-guided latent alignment'. In: *Pattern Recognition* 116 (2021), p. 107943. ISSN: 0031-3203. DOI: `https://doi.org/10.1016/j.patcog.2021.107943`. URL: `https://www.sciencedirect.com/science/article/pii/S0031320321001308`.

[69] Rui Wang et al. 'Instance Weighting for Neural Machine Translation Domain Adaptation'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1482–1488. DOI: `10.18653/v1/D17-1155`. URL: `https://aclanthology.org/D17-1155`.

[70] Xiaosong Wang et al. 'ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3462–3471. DOI: `10.1109/CVPR.2017.369`.

[71] Yichen Xie et al. 'Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 23715–23724.

[72] Wilson Yan et al. ‘Videogpt: Video generation using vq-vae and transformers’. In: *arXiv preprint arXiv:2104.10157* (2021).

[73] Donghyun Yoo et al. ‘Efficient k-shot learning with regularized deep networks’. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[74] Lei Zhang and Xinbo Gao. ‘Transfer Adaptation Learning: A Decade Survey’. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.1 (2024), pp. 23–44. DOI: 10.1109/TNNLS.2022.3183326.

[75] Tianjun Zhang et al. *Controllable Text-to-Image Generation with GPT-4*. 2023. arXiv: 2305.18583 [cs.CV].

[76] Kaiyang Zhou et al. ‘Domain Generalization: A Survey’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1–20. ISSN: 1939-3539. DOI: 10.1109/tpami.2022.3195549. URL: http://dx.doi.org/10.1109/TPAMI.2022.3195549.

List of Figures & Tables

Abbreviations & Terms

Chapter	Abbreviation/Term	Meaning	Pages
2	VAE	Variational autoencoders	10
	GAN	Generative Adversarial Networks	11
	GPT	Generative Pre-Trained Transformer	12
3	FTER	Fine-tune Efficiency Ratio	26
4	T-Gap	Trainability Gap	35