学位論文内容の要旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　李広

学位論文題名

A study on data-efficient learning and its medical applications
（データエフィシェントラーニングとその医療応用に関する研究）

Deep learning has experienced remarkable advancements and demonstrated remarkable achievements in various domains, such as computer vision, natural language processing, and speech recognition. In recent years, prominent deep learning models, including AlexNet, ResNet, BERT, ViT, CLIP, Stable Diffusion, and ChatGPT, have been developed and relied upon large-scale datasets for training. However, working with such large datasets poses significant challenges in terms of storage, transmission, and preprocessing. Additionally, training on large-scale datasets requires extensive computational resources, often involving thousands of GPU hours to achieve good performance. To address these challenges, the thesis focuses on investigating data-efficient learning methods.

Data-efficient learning is a subfield of machine learning that focuses on training models with limited amounts of data while maintaining high performance. Traditional machine learning algorithms often require large datasets to generalize well and make accurate predictions. However, in many real-world scenarios, collecting and labeling massive amounts of data can be time-consuming, expensive, or even impractical. Data-efficient learning aims to overcome these limitations and develop methods that can effectively learn from small or scarce datasets.

One method of data-efficient learning is transfer learning, where a pre-trained model on a large dataset is fine-tuned on a smaller target dataset. By leveraging the knowledge learned from the larger dataset, the model can quickly adapt to the new task with fewer training examples. This method has been successfully applied in various domains, including computer vision, natural language processing, and speech recognition. Another method of data-efficient learning is active learning, which involves selecting informative samples from a large pool of unlabeled data and actively querying human experts to label those samples. The labeled samples are then used to train a model, and the process iterates, gradually improving the model's performance with a minimal amount of labeled data. Active learning can significantly reduce the labeling effort and achieve good performance with a small labeled dataset. Furthermore, techniques such as semi-supervised learning and weakly supervised learning also contribute to data-efficient learning. In semi-supervised learning, models are trained using a combination of labeled and unlabeled data, where the unlabeled data provides additional information to improve the model's generalization. Weakly supervised learning, on the other hand, deals with tasks where only partial or noisy supervision is available, allowing models to learn from imperfect labels or weak annotations.

Data-efficient learning is a rapidly evolving field driven by the necessity to address real-world problems with limited data availability. While existing methods can alleviate some of the challenges posed by large-scale datasets, they inherently possess limitations when applied to certain scenarios. For instance,

constructing new datasets requires careful consideration of their complexity to effectively train neural networks. Moreover, existing methods may not be suitable for situations involving extremely limited data or labels. Therefore, there is a need for exploring stronger data-efficient learning methods to address these limitations.

The purpose of this thesis is to construct new datasets more efficiently and to enhance the learning capabilities of models when facing extremely limited data or labels. To achieve this goal, the thesis proposes a novel data-efficient learning method consisting of the following three stages. The first stage involves assessing the complexity of datasets by analyzing their characteristics and properties. Understanding the complexities of a dataset allows researchers to make well-informed choices regarding model architecture, training strategies, and data augmentation techniques that are appropriate for that particular dataset. This stage plays a crucial role in optimizing the learning process and achieving superior performance with limited data. Building upon the dataset complexity assessment, the second stage introduces the concept of dataset distillation. Dataset distillation leverages knowledge from a larger, labeled dataset to distill it into a smaller, more compact dataset. The distilled dataset retains the most relevant information that is essential for the target task. This stage can enhance data processing efficiency and avoid overfitting or noise from the large dataset. Lastly, the third stage explores self-supervised learning as a data-efficient learning method. Self-supervised learning involves training models to solve pretext tasks using unlabeled data, with labels generated automatically or through heuristics. The learned representations from these pretext tasks can then be transferred to the target task, effectively utilizing the large amounts of unlabeled data to improve performance. This stage can reduce reliance on labeled data while still achieving competitive results. With the incorporation of the three stages, the newly proposed data-efficient learning method can effectively address the existing challenges.

The structure of the thesis is shown as follows. Chapter 1 describes the research background and purpose of this paper. In Chapter 2, related works of data-efficient learning are presented and problems to be solved are clarified. In Chapter 3, a dataset complexity assessment method based on cumulative maximum scaled area under the Laplacian spectrum is presented. In Chapter 4, a method of generation of compressed gastric images based on soft-label dataset distillation for efficient anonymous medical data sharing is presented. In Chapter 5, a self-supervised learning method for learning discriminative representations from gastric X-ray images is presented. In Chapter 6, a self-supervised transfer learning method for COVID-19 detection from chest X-ray images is presented. In Chapter 7, for boosting COVID-19 detection accuracy, a novel method based on self-supervised learning and self-knowledge distillation is presented. In Chapter 8, the conclusions of the thesis and the future directions are discussed.

To summarize, the thesis introduces a new data-efficient learning method that encompasses three stages: dataset complexity assessment, dataset distillation, and self-supervised learning. This method aims to construct new datasets with improved efficiency and enhance model learning capabilities, particularly when dealing with severely limited data or labels. Additionally, the effectiveness of the proposed method is evaluated on both natural image datasets and medical image datasets.