Learning to Rank with Relative Annotation for Medical Image Analysis

門田, 健明

Learning to Rank with Relative Annotation for Medical Image Analysis

Takeaki Kadota

Department of Advanced Information Technology
Doctor of Engineering

KYUSHU UNIVERSITY

February 2023

Learning to Rank with Relative Annotation for Medical Image Analysis

Takeaki Kadota

Department of Advanced Information Technology
February 2023
Doctor of Engineering

Abstract

Deep learning has recently been expected to play a supporting role in disease severity level estimation of medical images. It achieves high performance by training on a large amount of labeled data. Absolute annotation, which represents continuous disease severity to discrete severity levels, is generally assumed to train deep learning for severity estimation. However, absolute annotation is difficult due to ambiguous level judgments. Consequently, attaching absolute annotations, even for medical experts, leads not only to high annotation costs but also to high variability in the judgment. In contrast, relative annotation, which represents which of two images has a higher severity, is expected to be simple and has low variability. This thesis proposes two methods for estimating disease severity by learning to rank with relative annotations. The first method is multi-task learning, which combines learning to rank by relative annotation and regression learning by absolute annotation. This method can substantially reduce the annotation cost by using relative annotations and a small number of absolute annotations. In a severity classification task of ulcerative-colitis endoscopic images, the proposed method achieved higher classification performance (accuracy and F1 score) with 10% annotation cost compared to the conventional classification methods only with absolute annotation. The second method is active learning using a Bayesian convolutional neural network (CNN) for determining image pairs to which relative annotations should be attached. This method is an active sample selection based on uncertainty given by the Bayesian CNN. The relative severity level estimation results showed that this method performs better than methods trained by randomly paired samples. The results also showed that this method is robust to class imbalance because it selects minor but important samples during its uncertainty-based selection.

Thesis Supervisor: Seiichi Uchida, Distinguished Professor, Kyushu University

Acknowledgments

I am deeply grateful to my supervisor, Professor Seiichi Uchida, for his guidance and for giving me the opportunity to become a member of his laboratory. In addition, I am deeply grateful to Professor Hideaki Hayashi for their support and advice over the years. I would also like to express my deepest gratitude to my committee members, Professors Atsushi Shimada and Ryoma Bise, for their discussion and supervision of this thesis. I am also grateful to the members of the Human Interface Laboratory for their discussions and advice on my research. Finally, I would like to thank my family for their understanding of my decision to pursue a doctoral degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Medical Image Analysis and Machine Learning

Recently, machine learning has been applied to various fields. In particular, with the development of deep learning, remarkable results have been achieved in computer vision tasks, such as image classification and regression, anomaly detection, segmentation, and image generation. Deep learning can extract highly non-linear and appropriate features from its high-dimensional inputs, such as images, by representation learning.

In medical image analysis, there has also been much research on the application of machine learning to support medical doctors. For example, there are studies on detecting shadows from chest X-ray images as an abnormality detection task [1, 2, 3] and detecting tumor cell regions from pathology images as a segmentation task [4, 5]. In addition, there are studies to identify disease severity and lesion differences for classification tasks [6, 7, 8]. In recent years, machine learning methods that perform as well as medical experts in these studies have been proposed. Therefore there is potential to reduce the burden on medical doctors in their clinical practice.

1.1.2 Relative Annotation for Medical Image Analysis

Deep learning requires a large amount of labeled image data for training, and therefore the cost of annotating medical images is an important practical issue. In fact, the annotation cost of medical image analysis tasks is much higher than those of general image analysis tasks. In general image analysis, crowd-sourcing services can efficiently create large amounts of annotation data. However, it is impossible to use these services for medical image analysis. This is because the skills possessed by medical doctors are required to detect the slightest change in condition without overlooking it.

A common annotation task for severity estimation is to attach a discrete severity level (such as 0, 1, 2, 3) to an image according to criteria defined by medical knowledge. Hereafter, I call this annotation *absolute annotation* because the absolute severity level is given as an annotation. Absolute annotation is difficult even for medical experts. The first reason for the difficulty is that it is difficult to find indicators that determine the severity of the disease from the image. It is necessary to carefully examine the entire image to find the subtle indicators. The second reason is that severity levels are inherently continuous, although the annotation labels are discrete. For example, the annotation task that attaches discrete severity levels 0, 1, 2, and 3 to lesions with continuous severities frequently finds images with intermediate severity levels, such as 1.5. Therefore, annotation results become inconsistent and unstable because of the divergence of judgments not only among multiple medical experts but also within the same medical expert.

A simpler annotation type is *relative annotation*, which compares the severity between two images. Specifically, relative annotation expresses the relationship between two images as to which is more severe. Relative annotation is appropriate for severity levels because it can deal with continuous levels; for example, even for two images A and B with severity levels 2.1 and 1.6, we can easily attach the relative annotation as "A is more severe than B." Moreover, relative annotation is expected to be more

consistent and stable because the annotator only has to choose one from two options ("A is more severe than B," or vice versa). According to these properties, relative annotation is a promising alternative to conventional absolute annotation.

Relative annotation can be directly utilized in a machine learning framework, called *learning to rank*, whereas absolute annotation has been mainly utilized for classification or regression. As detailed in Section 2.1, learning to rank is used for tasks to predict the order of samples by learning specific ordinal information from the ground truth. Learning to rank has been studied mainly in information retrieval as a task to predict the most important sentences when a query is given. In computer vision, classification and regression with absolute annotation have been extensively studied, but the application of learning to rank with relative annotation has not been fully explored. Therefore, this thesis mainly explores the effectiveness of learning to rank using relative annotations in medical image analysis.

This thesis also addresses relative annotation in class-imbalanced datasets. A class imbalance in medical image datasets is often an issue. The reason is that patients with severe diseases are generally fewer than those with mild diseases. Another reason is that the proportion of patient severity differs by the function and size of the hospital. Since relative annotation for class-imbalanced datasets has not yet been fully considered, a practical relative annotation mitigates class imbalance is considered.

1.2 Purposes

This thesis has two main purposes. Firstly, I propose a new method for severity estimation with relative annotation. Secondly, I propose a method to select sample pairs whose relative annotations are expected to be more important for the severity ranking task.

These two purposes are detailed as follows:

- **To explore automatic severity estimation with relative annotations:**

The application of deep learning to disease severity estimation from medical images requires a large amount of training data. However, the annotation cost of medical images is high because medical experts are required to annotate the images. This thesis explores learning to rank with relative annotation to reduce the annotation cost.

- **To investigate effective active learning for relative annotations:** In a learning-to-rank framework with relative annotation, the number of annotations for all possible pairs is still prohibitive. Therefore appropriate sample pair selection is important. This thesis explores an effective active learning method that automatically selects appropriate pairs for relative annotation. In addition, this thesis verifies the effectiveness of the proposed method on a dataset with class imbalance.

1.3 Contributions

Corresponding to the two purposes, the main contributions of this thesis are summarized as follows:

- **Proposing multi-task learning with learning-to-rank and regression with less annotation cost and higher classification performance:** I propose a multi-task learning method that combines learning to rank with regression for automatically estimating disease-severity levels. Using relative annotation, I demonstrate that the proposed method outperforms conventional multi-class classification methods and substantially reduces the annotation effort for medical image datasets.

- **Demonstrating that learning to-rank with Bayesian CNN could suppress the number of relative annotations:** I propose a deep Bayesian active-learning-to-rank that can rank the image samples according to their sever-

ity level. I demonstrated that the proposed method could suppress the number of relative annotations while keeping ranking performance. Experimental results confirm that the proposed method is robust against class imbalances by selecting minority but important samples.

1.4 Organization of Thesis

This thesis is organized as follows:

- **Chapter 1:** This chapter provides an overall background to this thesis. It outlines the purposes and contributions of this thesis.
- **Chapter 2:** This chapter describes an overview of preliminaries on learning to rank. First, the framework for learning to rank is briefly introduced, and the properties of the three learning-to-rank approaches are compared. Then, three annotations for learning to rank are introduced. Lastly, as a preliminary experiment, the performance evaluation of deep learning-to-rank using a logo images dataset is described.
- **Chapter 3:** This chapter describes an overview of medical image annotation for learning to rank. First, disease severity scores for medical image diagnosis are explained. Then, three medical image annotations for learning to rank are described.
- **Chapter 4:** This chapter describes the study of annotation cost reduction using relative annotation. The proposed method is explained, and the results of comparative experiments and an ablation study are provided.
- **Chapter 5:** This chapter explores a method that selects high-learning effect pairs for the relative annotation. First, the proposed active learning-to-rank is described. Then, experimental results and analysis of the proposed and comparative methods are provided.

- **Chapter 6:** This chapter concludes the thesis with future work.

Chapter 2

Preliminaries on Learning to Rank

2.1 General Framework of Learning to Rank

Learning to rank is an algorithm that automatically estimates the order of samples for ranking tasks. Figure 2-1 shows the learning-to-rank framework [9] consisting of training and testing steps. The training step is to train a ranking model by training data. First, input samples are attached with ordinal information as ground truth by annotation. Second, the ranking model is optimized from the learning-to-rank algorithm using samples with the ground truth as training data. The testing step is to sort test data using the trained ranking model. Many learning-to-rank algorithms have been proposed, and they can be categorized into three approaches [9], as we will see in Section 2.2. In addition, three annotation methods for learning to rank have also been proposed, as we will see in Section 2.5.

2.2 Three Approaches of Learning to Rank

Learning-to-rank algorithms can be categorized into three main approaches (pointwise, pairwise, and listwise approaches). It is necessary to use different approaches depending on the properties of the task (e.g., total ranking, top-k ranking, and bipar-

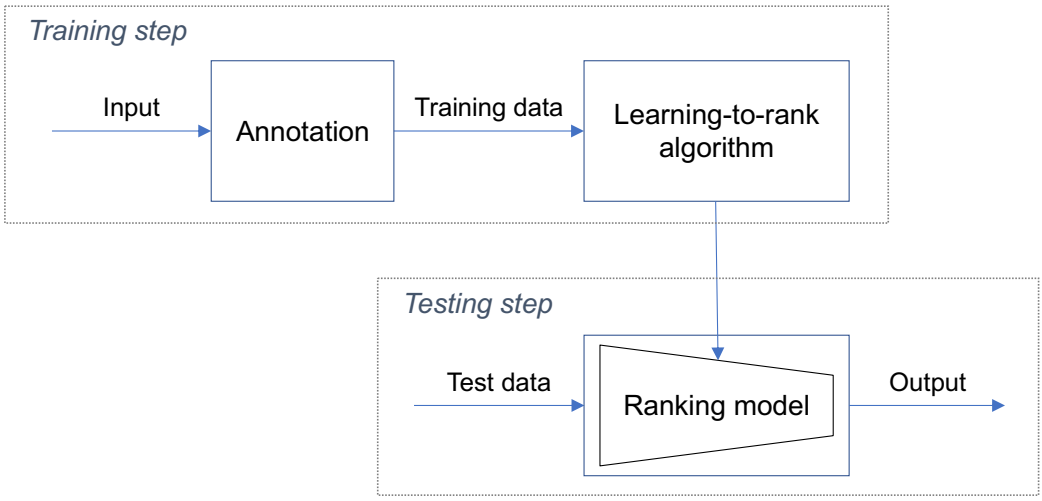


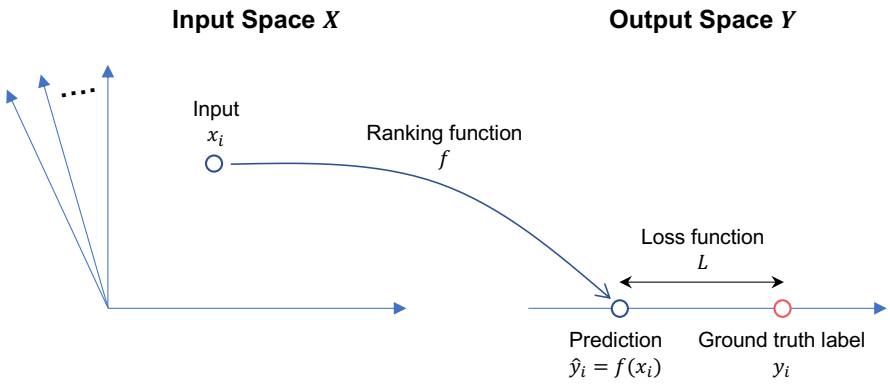
Figure 2-1: General framework of learning to rank.

tite ranking) and the dataset (e.g., text data, audio data, and image data). In order to distinguish between these three approaches, I summarize the properties of these approaches with deep learning.

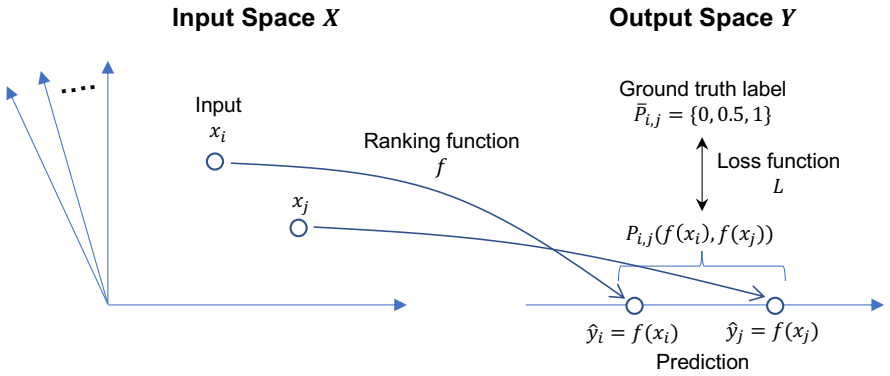
The differences in the properties of the three approaches can be explained by focusing on the set of input samples (pointwise, pairwise, and listwise sets), ground truth, and loss function. Figure 2-2 shows typical three deep learning-to-rank approaches. This figure is inspired by Liu [9] and represents the approaches by the four components (input space, output space, ranking function, and loss function). These approaches differ mainly in the type of input sample set used in the learning algorithm. An appropriate ground truth and loss function should be used depending on the input sample sets.

The pointwise approach

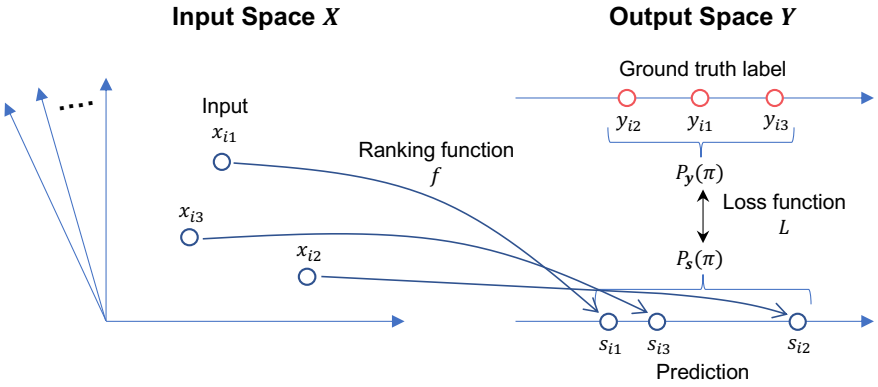
The input of the pointwise approach is a single sample, with one ground truth added to each sample. Each sample is attached with categorical, ordinal, or quantitative information as ground truth. The pointwise approach uses regression or classification methods for a ranking task. Therefore, the loss function uses a function in regression or classification methods, such as mean squared error (MSE) and cross-entropy loss



(a) Pointwise approach (Regression)



(b) Pairwise approach (RankNet)



(c) Listwise approach (ListNet)

Figure 2-2: Typical three deep learning-to-rank approaches.

functions. Many pointwise approaches have been proposed, such as PRanking [10], Logistic Regression-based Ranking [11], and SVM-based Ranking [12].

Figure 2-2(a) shows a regression framework for learning to rank. Let $f(\cdot)$ be a ranking function to rank samples. Given a set of samples $\mathbf{x} = \{x_i\}_{i=1}^N$ where N is the number of samples and ground truth labels $\mathbf{y} = \{y_i\}_{i=1}^N$, $y_i \in R$, the squared-error loss function L_{reg} is defined as follows:

$$L_{\text{reg}}^i(f; x_i, y_i) = (f(x_i) - y_i)^2. \tag{2.1}$$

The pairwise approach

The inputs of the pairwise approach are two samples, with one ground truth added to each sample pair. The ground truth is the ordinal information between two samples. In the pairwise approach, the ranking model can be considered a classification model of pairs, and the loss function uses mainly cross entropy as a classification function. The pairwise approach includes RankingSVM [13], RankNet [14] and SortNet [15]. This thesis focuses on RankNet, to which neural network architecture is applicable.

Figure 2-2(b) shows a framework of RankNet. In RankNet, given a set of sample pairs $\mathcal{D} = \{x_i, x_j, y_{i,j}\}$, $i, j \in [1, N]$ where ground truth $y_{i,j}$ is a relative label of the pair (x_i, x_j), suppose a ranking function $f(x)$ to rank samples. Ground truth $y_{i,j}$ takes one of three values $\{0, 0.5, 1\}$ for small, equal, or large by comparing a sample pair and is regarded as a target probability $\bar{P}_{i,j}$. Then, let $f(x_i)$ and $f(x_j)$ be the rank scores of the sample pair, and $o_{i,j} = f(x_i) - f(x_j)$ be their difference, a probability $P_{i,j}$ is defined as follows:

$$P_{i,j}(f(x_i), f(x_j)) = \frac{\exp(o_{i,j})}{1 + \exp(o_{i,j})}. \tag{2.2}$$

Then, the cross-entropy loss function L_{rank} is defined from the target probability

$\bar{P}_{i,j}$ and the probability $P_{i,j}$ obtained from ranking scores as follows:

$$L^{i,j}_{\text{rank}}(f; x_i, x_j, y_{i,j}) = -\bar{P}_{i,j} \log P_{i,j} - (1 - \bar{P}_{i,j}) \log(1 - P_{i,j}). \tag{2.3}$$

The listwise approach

The input of the listwise approach is a sample list, with ordinal ground truth added to the list. The listwise approach includes AdaRank [16], ListNet [17], and ListMLE [18]. Here I describe ListNet, which can be applied to deep learning.

Figure 2-2(c) shows a framework of ListNet. In ListNet, given list set $\mathbf{x} = \{x_i\}_{i=1}^N$ and ground truth labels $\mathbf{y} = \{y_i\}_{i=1}^N$ which is total order labels of the list, suppose a ranking function $f(x)$ to rank samples. Furthermore, given a ranking score list $\mathbf{s} = \{s_i\}_{i=1}^N$ obtained from $f(x)$, the permutation probability $P_{\mathbf{s}}$ is defined from the Plackett-Luce model as follows:

$$P_{\mathbf{s}}(\pi) = \prod_{j=1}^{N} \frac{\varphi(s_{\pi(j)})}{\sum_{k=j}^{N} \varphi(s_{\pi(k)})}, \tag{2.4}$$

where π is a permutation of samples, $s_{\pi(j)}$ the ranking score of the sample at jth position of permutation π, and φ is a monotonically-increasing function. The loss function L_{list} is the Kullback-Leibler divergence loss between the ranking-score permutation probability $P_{\mathbf{s}}(\pi)$ and the ground-truth permutation probability $P_{\mathbf{y}}(\pi)$, and thus is defined as follows:

$$L_{\text{list}}(f; \mathbf{x}, \mathbf{y}) = D(P_{\mathbf{y}}(\pi) || P_{\mathbf{s}}(\pi)). \tag{2.5}$$

2.3 Comparison of Learning-to-Rank Approaches

Table 2.1 compares the three learning-to-rank approaches. "Order relation" is the strength of using ordinal information between samples for training. "Task trans-

Table 2.1: Properties of three learning-to-rank approaches.

Approaches	Order relation	Task transformation
Pointwise	low	yes
Pairwise	high	no
Listwise	very high	no

formation" indicates the possibility of adapting for other tasks (e.g., regression or classification) beyond the ranking task.

The advantage of the pointwise approach is that it can be treated as classification or regression tasks as well as ranking tasks. Since the pointwise approach predicts scores for individual samples, standard classification and regression are used as ranking tasks. In addition, the pointwise approach can be applied to ordinal regression tasks; for example, regression is used as a naive method [19].

The disadvantage of the pointwise approach is that it cannot directly reflect the order relations among samples as training information. For ranking tasks, accurate order estimation among samples is important. Therefore, the pointwise approach is not actively selected for the ranking task compared to the pairwise and listwise approaches, which can directly learn ordinal information among samples.

The advantage of the pairwise approach is that it can directly reflect the order relations among samples to the ranking model by the ground truth of the order relations of the paired data. Since the pairwise approach can utilize more ordinal information than the pointwise approach, it is a suitable learning-to-rank method for ranking tasks using datasets attached with relative labels.

The disadvantage of the pairwise approach is that the output values are only relative because it uses ground truth as the ordinal relation for two samples. Therefore, the results can only be evaluated with output values from multiple samples. The pairwise approach can be used only for ranking tasks and not for regression or classification tasks.

The advantage of the listwise approach is that it can directly reflect the order rela-

tions among multiple samples for training using ground truth in the ordinal relations in the list. The listwise approach is suitable learning to rank for ranking tasks using datasets attached with list order.

The disadvantage of the listwise approach is that, like the pairwise approach, the output values do not represent absolute scores. Therefore, the listwise approach can only be used for ranking tasks, not regression or classification tasks.

This thesis primarily employs a pairwise approach, which can directly learn ordinal information between samples. The training data obtained by relative annotation, which is the focus of this thesis, is more suitable for use in the pairwise approach than in the pointwise or listwise approach, as will be discussed later in Section 3.3 of Chapter 3.

2.4 Related Work of Learning to Rank

Learning to rank has been widely used in recommendation systems in the past. In recent years, it has been applied to several image analysis problems. For example, in image quality evaluation and image attractiveness evaluation, the ranking function is applied because it is difficult to give an absolute quality evaluation for each image [20, 21, 22, 23, 24, 25, 26].

Some studies have also reported the application of learning to rank for medical image analysis [27, 28, 29, 30]. The studies by Huang et al., Pedregosa et al., and Peng et al. did not use deep learning-to-rank but instead learning to rank by simple or handcrafted features [27, 28, 29]. Lyu et al. apply deep learning-to-rank to select high-quality images in ultrasound images [30]. Their study did not use learning to rank for the disease severity estimation task. Therefore it is not a study that verifies the effectiveness of deep learning-to-rank for medical image diagnosis.

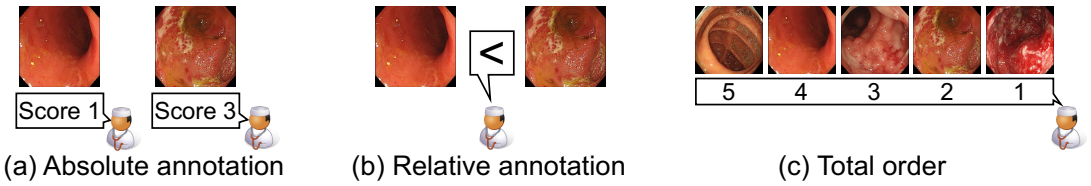


Figure 2-3: Three annotation methods for learning to rank.

2.5 Three Annotation Methods for Learning to Rank

There are three annotation methods for learning to rank. Figure 2-3 shows three annotations for learning to rank. Note that these annotation methods are theoretically applicable to arbitrary learning-to-rank approaches, but in practice, they are closely related to the three approaches to learning to rank. The details are described in Section 2.6.

According to Liu [9], three annotation methods are relevance degree, pairwise preference, and total order. Since this thesis focuses on applying deep learning to image analysis, we rename them for image annotation. Specifically, relevance degree is called "absolute annotation," and pairwise preference is called "relative annotation".

The absolute annotation

The absolute annotation is a standard method that attaches a categorical or quantitative label to a sample. Given a sample x_i, an absolute label y_i is attached to the sample. The annotated sample set is represented as $\mathcal{D}_{\mathrm{L}} = \{(x_i, y_i)\}$. For N samples, the total number of annotations is N. The amount of ordinal information in a dataset depends on the number of categories (classes) attached to the sample.

The relative annotation

The relative annotation is a method that attaches the ground truth label of the ordering relationship between two samples. For a sample pair (x_i, x_j), a relative label

$y_{i,j}$ is regarded as a target probability $P_{i,j}$ and is defined as follows:

$$P_{i,j} = \begin{cases} 1, & \text{if } x_i \text{ has a higher level than } x_j, \\ 0.5, & \text{else if } x_i \text{ and } x_j \text{ have the same level,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

The set of the annotated sample pairs is represented as $\mathcal{D}_{\mathrm{L}} = \{(x_i, x_j, P_{i,j})\}$. In the relative annotation, the total number of annotations increases quadratically with the number of samples. For N samples, the total number of annotation pairs is $N(N-1)/2$. Therefore, annotating all pairs becomes more difficult as N increases. The relative annotation can give more ordinal information than the absolute annotation by comparing samples in the same class.

The total order

The total order is an annotation that attaches ordinal information to the entire dataset. Given a sample x_i, a total order label y_i is attached to the sample. The annotated sample set is represented as $\mathcal{D}_{\mathrm{L}} = \{(x_i, y_i)\}$. For N samples, the total number of annotations is N. However, the annotation becomes more difficult as the number of samples increases because of the need to compare the ordinal relationship of all samples. The total order can give the dataset more ordinal information than absolute or relative annotations.

2.6 Relationship between Learning to Rank and Annotation

The relationship between the three learning-to-rank approaches and the three annotation methods is described below. Figure 2-4 shows the relationship between three annotations and three learning-to-rank approaches. As mentioned above, note that learning-to-rank approaches and annotation methods are independent of each other.

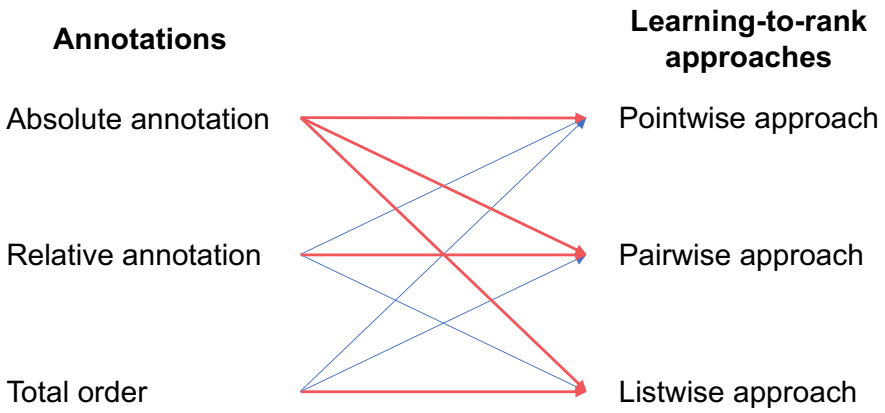


Figure 2-4: Relationship between three annotations and three learning-to-rank approaches. Red arrows indicate approaches that enable learning to take advantage of ground truths' properties.

The three learning-to-rank approaches can apply the ground truth obtained by the three annotation methods. An appropriate learning-to-rank approach should be selected based on the annotation property and target task. The following describes how to use the ground truth obtained by each annotation for learning to rank.

First, in absolute annotations, each sample in a dataset is attached with one ground truth, and thus the dataset can be used for a pointwise approach. The dataset can be adapted to the pairwise approach by creating pairs of samples and reattaching the ground truth between pairs. In addition, the dataset can be used for a listwise approach by using the absolute-label order as a list.

Second, in relative annotations, each paired sample is attached to a ground truth about the ordinal information, and thus the dataset can be used in a pairwise approach. Furthermore, the dataset can be applied to the listwise approach by transforming the order of the dataset from pairs to lists. The sample order in a list can be used for a pointwise approach by considering it as the ground truth of each sample. However, pair-to-list order transformations generally result in incomplete information because of the large amount of missing pair information. Therefore, relative annotations are best applied to the pairwise approach.

Finally, in total order, each sample is attached to a ground truth about the order in the entire dataset, and thus the dataset can be used for a listwise approach. By pair creation from a list, the dataset can be adapted to the pairwise approach. Furthermore, the dataset can be adapted to a pointwise approach by using the order number of the list. However, total order has much ordinal information, and it should not be used for pairwise or pointwise approaches that use only some of the ordinal information. The total order is best applied to a listwise approach.

2.7 A Basic Experimental Study of Deep Learning-to-Rank

2.7.1 Representation Learning and Deep Learning-to-Rank

Recently, there has been much research on using machine learning to replace tasks performed by humans. Machine learning is a technique that automatically processes a target task (e.g., classification or regression) by extracting patterns from data. The efficient extraction of patterns is performed by representing the data as appropriate information (called *feature*). In conventional machine learning, the representation method was designed by a human, but recently an approach that learns the representation itself (called *representation learning*) has been proposed. Representation learning reduces human intervention and automatically provides features to extract patterns effectively.

Deep learning is a representation learning that can automatically extract abstract features from data. The deep learning model is a deep multi-layer perceptron, a type of neural network, a mathematical function that maps input values to output values. Deep learning has achieved high estimation performance for various complex tasks.

Deep learning can be applied to RankNet, which has a neural network structure. On the other hand, RankNet is designed with shallow multi-layer perception for

information retrieval and has not been fully explored with deep neural networks for image analysis. Therefore, it is necessary to verify that deep learning-to-rank can effectively extract features as representation learning in image analysis.

2.7.2 Logo Ranking with Deep Learning-to-Rank

This research studies the effectiveness of CNN-based representation learning for learning to rank in a relative popularity estimation task from logo images, a task for which abstract features are challenging to extract. A logo is a graphic design widely used to identify a company, organization, group, or product. A logo mark must capture the subject's characteristics and minimize waste to be easily identifiable and memorable. Therefore, professional designers carefully create each logo with various aspects considered. Such a logo mark is not merely a symbol but expresses the visual meaning of the company or organization's policy, history, philosophy, or commercial strategy.

Logos can be visually categorized into three types in general: logotypes, logo symbols, and mixed types. Figure 2-5 shows examples of each type of logo. A em Logotype is a logo with no picture and consists only of letters, often representing a company's name or initials. *Logo symbol* is a logo composed of an abstract mark, icon, or pictogram. *Mixed* is a mixed logotype and logo symbol containing both textual and pictorial information. There are several other variations of logotype classification. For example, in [31], three types are "text-based logo", "iconic or symbolic logo", and "mixed logo".

There has been much research on the design of logos and their impressions. Zhao [32] discusses the recent diversity of logos in the survey paper. Sundar et al. [33] analyzed the impact of the logo's vertical position on the package. Luffarelli et al. [34] used hundreds of crowd workers to analyze the difference between symmetry and asymmetry in logo design. They concluded that asymmetrical logos make more *arousing* impressions.

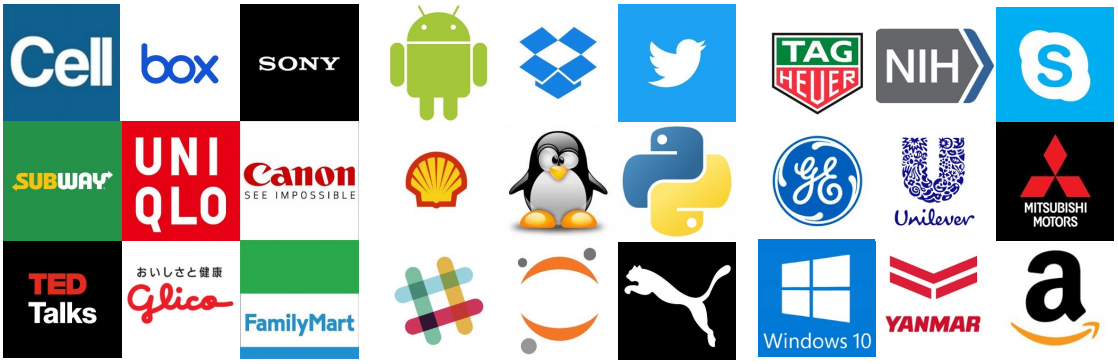
Luffarelli et al. [35] recently reported the exciting result that a logo's *descrip-*

tiveness positively impacts brand reputation. The descriptiveness means that the text and illustration design of the logo explicitly expresses what the company is doing. This report provides a subjective analysis using crowd workers and discusses the importance of text in logo design.

In computer vision, research dealing with logo design has recently been conducted, primarily because of the release of the public logo image dataset. The early dataset, such as the UMD-Logo-Database (currently unavailable), had only 123 logo images and was small in size [36]. By contrast, the size of the logo dataset is now considerably more significant. WebLogo-2M [37] is the earliest large logo dataset containing 194 different logos captured in 1,867,177 web images. They apply the dataset for the logo detection task. Sage et al. [38] released the LLD-logo and the LLD-icons datasets. The LLD-logo dataset contains 122,920 logo images collected from Twitter and 486,377 favicon images (32×32 pixels) from web crawling. They use these logo images for the logo synthesis task. In addition, a Logo-2k+ dataset containing 2,341 different logos in 167,140 images was recently published by Wang et al. [39]. The unique feature of this dataset is that the logo images collected by web crawling are classified into ten different company classes (e.g., food, clothing, institutions, and accessories). They use these logo images for the logo classification task.

Although logo research has been conducted in various tasks, it has yet to be common to conduct research on logo designs for objective and quantitative evaluation. In an exceptional attempt, Karamatsu et al. [40] explore the analysis of favicons in LLD-icons using a top-rank learning method to identify trends in favicon design for each company type. This research attempts to quantify logo design by analyzing the relationship between logo images and their popularity (given the number of followers on Twitter).

The primary purpose of this research is to analyze the relationship between the visual appearance of logo images and the number of followers of a company or organization. For this purpose, I utilize state-of-the-art pairwise learning-to-rank and a



(a) Logotype. (b) Logo symbol. (c) Mixed.

Figure 2-5: Logo image examples.

large logo dataset, LLD-logo [38]. Specifically, RankNet [14] (originally a multilayer perceptron) is used in convolutional neural networks. RankNet has many applications, such as image attractiveness evaluation [23]. Recent applications of related techniques about RankNet are well summarized by Guo et al. [41]. To the best of the authors' knowledge, an attempt has yet to be made in recent years to analyze logo images using objective quantitative evaluation using machine learning. Through the analysis of logos and the number of company followers, trends in logo design used by companies and organizations can be identified. Furthermore, the analysis results can help design new logos that match the company's intentions.

The main analysis is the relationship between the overall logo image (including text and symbols) and the number of company followers with the logos, using learning to rank and regression. In this research, the number of company followers is considered to be an indicator of the company's popularity. If it is possible to predict the number of followers from logo images with a reasonably high accuracy using ranking and regression functions, these functions would be influential in determining whether the logo design is commensurate with the company's popularity when logo creation.

This research uses the LLD-logo dataset by Sage et al. [38]. LLD-logo contains 122,920 logo images, which are collected from Twitter profile images using an API called `Tweepy`. This dataset has undergone careful filtering operations on the collec-

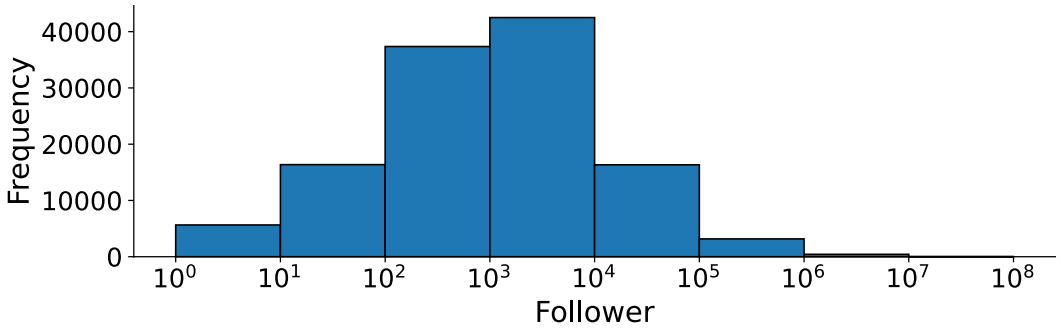


Figure 2-6: Distribution of followers of the companies listed in LLD-logo dataset.

tion to exclude facial images, harmful images, and illustrations. The size of the logo image included in this dataset is $63 \times 63 \sim 400 \times 400$ pixels.

The logo images in the LLD-logo dataset have two advantages over images taken with cameras like the Logo-2K [39]. First, because LLD-logo images are so-called born-digital, they are not affected by disturbances such as uneven light, geometric distortion, low resolution, or blur. Second, an advantage of LLD-logo is metadata, of which *the number of followers* is an excellent indicator for understanding the popularity of individual companies. Figure 2-6 shows the distribution of the number of company followers in the LLD-logo dataset. Note that the horizontal axis of this histogram is expressed in logarithmic form. Such a log transformation is a common method used to analyze popularity, especially Twitter followers distribution [42, 43, 44, 45].

The main contributions of this research are summarized as follows:

- To the best of our knowledge, this research is the first large-scale objective analysis of the relationship between logo image and a company's popularity (number of followers).

- Ranking and regression analysis indicated the potential for estimating relative popularity among logo images and absolute popularity (number of followers) from logo images. The results suggest that the learning-to-rank and regression functions can be used to evaluate the quality of a logo design. These functions may become clue tools for logo design strategies in logo production.

2.7.3 Estimating Relative Popularity through Logos

I explored the feature extract effectiveness of deep learning-to-rank from the experiment of the relative popularity estimation through logo images. The number of followers depends on many factors, including the company's products, market trends, size, and impressions. In addition, different people have different reasons for following a company. Therefore, estimating the relative popularity between logo images is generally considered difficult. In this research, I attempted to estimate the relative popularity between logo images by extracting highly abstract concepts contained in the logo image using deep learning, which has high expressive performance.

More specifically, I conduct experiments on the relative estimation of the number of company followers as a ranking by combining deep representation learning and RankNet [14]. RankNet is a pairwise learning-to-rank approach, which is to determine a ranking function $r(x)$ that satisfies the condition $r(x_1) > r(x_2)$ if x_1 should rank higher than x_2 in the two samples. In this case, x represents the logo image, and if the logo image x_1 has more followers than x_2, this condition should be satisfied. Given a ranking function $r(x)$ that outputs accurate rank scores $r(x)$, it is also possible to predict *relative* popularity, such as $r(x_1) > r(x_2)$ or $r(x_1) < r(x_2)$ for a given image pair x_1, x_2.

RankNet is a learning-to-rank method with a neural network framework and realizes a nonlinear ranking function. The original RankNet is implemented in shallow multi-layer perception (MLP). In this trial, I implemented DenseNet-169 to realize more substantial nonlinear functions and deal with image input to RankNet.

All logo images were resized to 224×224 pixels. The training, validation, and testing datasets are 30,000, 10,000, and 20,000 randomly selected logo images from the LLD-logo dataset, respectively. Then, I randomly create 30,000, 10,000, and 20,000 "pairs" from these data sets and use them for training, validation, and testing of RankNet with DenseNet-169. For training with RankNet, the cross-entropy loss was used. The training was terminated if the validation accuracy did not improve for

ten epochs.

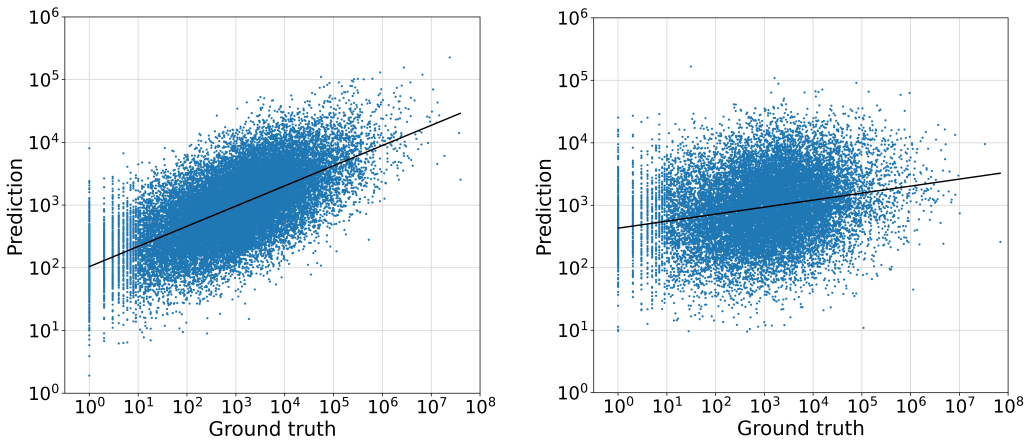
The performance evaluation of the learning-to-rank task differs from that of the regression task. Learning-to-rank task is evaluated by the correctness rate of the large/small comparison between the rank scores $\{r(x_1), r(x_2)\}$ calculated from the test pair $\{x_1, x_2\}$. When x_1 has a higher number of followers than x_2, the relative estimation is considered successful if the learned RankNet correctly evaluates $r(x_1) > r(x_2)$. The percentage of correct answers for the pair is 50% when the decision is made randomly.

The performance evaluation on the ranking task was 60.13% and 57.19% correct for the 30,000 training and 20,000 testing pairs, respectively. Since the estimation task of the follower number from logo images is generally difficult, these results, in which RankNet was used to determine the superiority between two logo images with about 60% accuracy, are considered to have higher performances than expected. The results of the learning-to-rank task suggest that the logo image and the number of followers contain elements that provide a correlation. Therefore, from the experiment of estimating relative popularity from logo images, I found that deep learning-to-rank can effectively extract features as representation learning in image analysis.

2.7.4 Additional Experiments by Regression

I estimate the absolute popularity from a logo image by regression, as another estimation task different from the relative popularity estimation by learning to rank. Regression estimates the absolute number of followers for the input logo images with deep learning. The model of the CNN backbone is used DenseNet-169. The regression loss function uses standard mean squared error (MSE). The same logo images in the experiment of learning-to-rank are used as the training, validation, and test sets. For the same reasons as in Section 2.7.2, I also used log-transformed the number of followers during training and testing.

Figures. 2-7 (a) and (b) are scatter plots of ground truth (GT) and prediction



(a) Training samples. (b) Test samples.

Figure 2-7: Prediction result of the number of followers from logo images.

result (PR) for the training and test samples, respectively. The black lines are the linear regressions obtained from GT and PR. The linear regression results show that GT and PR are weakly correlated, which means that the number of followers can be predicted to some extent (at least not impossible). For the training and test sets, the Pearson correlation coefficients for GT and PR are approximately 0.64 and 0.23, respectively. The test set correlation is not very strong, but the training set confirms that GT and PR are correlated. Also, p of correlation coefficients for the training and test sets were almost zero. In addition to the ranking task, experimental results from the regression task also showed that the number of followers can be estimated with good performance to some extent.

Furthermore, the accuracy of the relative popularity was determined using the absolute value of the number of followers estimated by the regression, with the same dataset as the image pairs used in the ranking task. The regression obtained the accuracy of the relative popularity of 71.9% and 57.8% for the training and test pairs, respectively. RankNet had almost the same relative popularity in test pairs compared to the regression. This result indicates that deep learning-to-rank has the same effective feature extraction as conventional deep regression in ranking tasks.

2.7.5 Conclusion

In this research, to confirm the effectiveness of applying deep learning to learning-to-rank, I explored the correlation between logo images and the number of Twitter followers in logo image analysis. Since the experimental results of the learning-to-rank task showed some follower estimation performance, I confirmed the effectiveness of implementing a learning-to-rank framework with CNNs. Thus, the results of the learning-to-rank task indicate that the logo image and the number of followers contain factors that give a correlation. Therefore, I found that deep learning-to-rank could effectively extract features in image analysis from the experiment of relative popularity estimation from logo images.

Chapter 3

Medical Image Annotation for Learning to Rank

3.1 Disease Severity Score for Medical Image Diagnosis

Annotation in medical image analysis is closely related to the medical image diagnosis by medical experts in clinical practice. Here, we pick up disease severity scores as a typical annotation in medical image diagnosis. The disease severity scoring task is mainly an ordinal classification task by medical experts. For example, the Mayo Endoscopic Score (MES), which is a severity score for ulcerative colitis (UC), is defined as normal, mild, moderate, and severe. According to Schroeder et al. [46], these categories have an ordinal relationship. The severity score is a discrete value, whereas the severity of the disease varies continuously. Therefore, medical experts determine each severity score by looking for landmarks that can be used as indicators of the score. For example, if there is erosion, the score is determined as Mayo 2. The severity score is important information for determining the treatment plan and understanding the disease's status.

These medical imaging severity scores are valuable indicators but also have problems. The main problem is that detailed information is lost when discrete scores represent disease severity. The discrete scores cannot represent detailed quantitative information. Discrete scores may be expressed as scores for apparent changes (e.g., the intensity of erythema, amount of bleeding), but they are qualitative evaluations from images, not quantitative evaluations. In addition, because multiple landmarks often determine the score, it does not represent an index for a specific finding, such as bleeding or ulcer size. Therefore, medical experts do not make treatment decisions based only on severity scores but also on detailed information from imaging findings. Therefore, many discrete severity scores used in clinical practice are important indicators for consistent treatment, but they also do not accurately represent the disease's original severity.

3.2 Three Annotation Methods for Medical Images

Three annotation methods for learning to rank are described in medical image analysis. The figure 2-3 of three annotations for learning to rank in Section 2.5 shows an example of severity annotation for endoscopic images of ulcerative colitis. The ground truth for learning to rank is attached to a medical dataset by medical experts.

The absolute annotation for medical images

The absolute annotation is a standard method that attaches a single label to a medical image. A severity score is attached to an image as a label in the disease severity estimation task. In medical image analysis, the annotation uses medically defined scoring for the target disease, which is the score that medical experts generally use in their practice. Therefore, because the labeling is used to determine the treatment strategy accurately, the annotator requires sufficient knowledge and experience with the target disease. In addition, as mentioned in Chapter 1, it is difficult for medical experts to accurately attach absolute labels to medical images because the continuous

disease severity is attached as a discrete severity by absolute annotation. As long as absolute annotation is used to create datasets for deep learning, the issues of annotation cost and labeling errors are invariably encountered.

The relative annotation for medical images

The relative annotation attaches the ground truth label of the ordering relationship between two medical images. The main ordering criterion is the relative disease severity in medical image analysis. In clinical practice, medical experts compare a given point in time of the disease to past conditions to determine changes over time. This is a similar task to determining relative annotations. For example, in a patient with a disease that repeats inflammation and remission, the change in the disease state is checked by comparing the images taken as a follow-up observation. In addition, changes between images taken before and after medication are observed to determine the effectiveness of the treatment. Relative annotation is much easier than absolute annotation because the annotator does not have to identify discrete severity levels for difficult samples with intermediate severity, such as level 1.5.

The total order for medical images

The total order is an annotation that attaches ordinal information to the entire medical image dataset. For example, this annotation is an operation to sort 10000 images of a disease in order of severity. The total order is an operation that medical experts rarely have the opportunity to perform in clinical practice. Because it is necessary to judge slight differences in severity, The total order should be performed by skilled medical experts with abundant clinical experience.

Table 3.1: Comparison of three annotations for learning to rank in medical image analysis.

Annotations	Medical definition	Ordinal information	Annotation cost	Imbalance impact
Absolute	yes	low	high	no
Relative	no	high	low	yes
Total order	no	very high	very high	no

3.3 Comparison of Annotation Methods for Medical Images

In this section, I compare the three annotation methods in medical image analysis. Table 3.1 shows the comparison of the three annotations for learning to rank in medical image analysis. "Medical definition" indicates that the scoring conditions are defined as medical findings. "yes" means that the decision is based on medical findings, and "no" means that the decision is based on the medical expert's intuition. "Ordinal information" represents the amount of ordinal information contained within the dataset. "Annotation cost" represents the time required to annotate an image, pair, or list once. Note that the annotation cost depends on the skill level of the medical expert, so it is assumed that an experienced medical expert performs the annotation. "Imbalance impact" represents whether an imbalance increases by annotation if the dataset has an imbalance.

The advantage of absolute annotation is that labels are clearly defined as medical standards, such as disease classification and severity. For example, the MES at UC defined erosion, ulcer, and so on as landmarks for judgment. Because these are defined as medical findings, the validity of the labeling judgment is medically guaranteed.

The disadvantage of absolute annotation is the high annotation cost for detailed annotation. There are three main reasons for the high annotation cost. First, the labeling criteria are based on medical standards, so annotations must be performed by medical experts who have sufficient knowledge of the disease. The second is that it

takes time to find the landmarks because they must be found in the image. Third, a disadvantage of absolute annotation is the low reproducibility of annotations because continuous changes in lesions are converted into discrete scores. For example, many intermediate lesions between scores 1 and 2 are found, such as score 1.5. Thus absolute annotation can lead to variability in results.

The first advantage of relative annotation is that comparing two images is easy, and the annotation cost is low for one pair. For example, comparing the severity of two images, the image with more collapsed tissue can be intuitively judged as severe. The images can be considered equal in the case of similar disease severity. Thus, the labeling decision can immediately be made because there is no need to search for landmarks as severity indicators. Second, relative annotations can create a dataset with more ordinal information than absolute annotations. Relative annotation allows ordinal information according to differences in severity, even between images attached to the same class in the absolute annotation.

There are two main disadvantages of relative annotation. First, the number of annotations quadratically increases as pairs are created. $N(N-1)/2$ image pairs can be created from N images. For example, about 50 million pairs can be created from 10,000 images. Therefore, it is practically impossible to annotate all pairs created from the dataset in relative annotations. Second, pair creation increases the influence on the imbalance of the dataset. When the imbalance is significant, most pairs are created from major classes. In addition, because the class imbalance cannot be known in the relative annotation, class imbalance mitigation cannot be performed before the pair is created, such as re-sampling.

The advantage of the total order is that it can attach the most ordinal information to the dataset of all the annotation methods for learning to rank. Therefore, ranking models can achieve high predictive performance by using the dataset attached to the total order for learning to rank.

Datasets with total order have much ordinal information but also many disadvan-

tages. There are two main disadvantages to the total order. The first is that the annotation cost is very high. For example, it is a very burdensome task to order 10,000 medical images in severity order by medical experts. Second, the reproducibility of the overall order is low. The overall sample order of the dataset is never the same among medical experts. Therefore, there is no way to guarantee the ordinal quality of the dataset other than to have the dataset created by experienced medical experts.

Relative annotation is the most suitable for medical image annotation among these three annotation methods. Relative annotation is easy to annotate a pair of images, and there is low variability in the annotation results. Relative annotation has the potential to reduce annotation costs in medical image dataset creation. Therefore, this thesis focuses on relative annotation and explores learning to rank for medical image analysis. In addition, this thesis also investigates the issue of increasing the number of pairs in relative annotation and the selection of effective pairs for class-imbalanced datasets.

3.4 Related Work of Relative Annotation

Some studies on learning to rank with relative annotation have been reported in image analysis. The ranking task is often converted to age estimation task [47, 48, 49] and medical analysis task [50, 51, 52, 53] based on a previous study [54]. These studies use binary classification to determine if the input sample is larger than a certain level. These studies cannot take advantage of relative annotations, which I focus on because they require absolute annotations. Learning to rank with relative annotation is used to predict relative attributes, which indicates the strength of attributes between images [55, 56, 57, 58]. These studies estimate visual relative attributes to general images by learning to rank, not applying them to medical image analysis.

Several studies using relative annotation in medical image analysis have been

reported [59, 60]. Kalpathy-Cramer et al. [59] showed, using a fundus image dataset of retinopathy of prematurity that absolute annotation that absolute annotations have high variability among annotators, whereas relative annotations have a high consistency among annotators. Furthermore, their study points out that ranking scores estimated using relative annotations can be helpful as continuous severity scores for diseases in clinical practice. However, their study did not apply deep learning to learning to rank and did not demonstrate sufficient performance in estimating disease severity. Their trained model only can estimate ranking scores and cannot be used for disease severity classification.

Saibro et al. [60] also demonstrated that relative annotation has a lower label error rate than absolute annotation using ultrasound images of nonalcoholic fatty liver disease. In addition, their study examined transforming a ranking task into a binary classification task by determining a threshold on the ranking score by medical experts. However, although their study uses deep learning-to-rank, their method can not be applied to multi-class classification. Furthermore, the ranking score threshold for binary classification is subjective; thus, the classification performance can vary by the annotator.

The deep learning-to-rank with relative annotation examined in this thesis differs from their studies in three main respects. First, I focus on reducing annotation costs as an advantage of relative annotation. Second, I adapt learning to rank with relative annotation to multi-class classification using a method that calibrates ranking scores to disease severity scores. Third, I explore an active learning method that selects high-learning effectiveness pairs to address the increased number of pairs that can be annotated, which is an essential issue for relative annotation.

Chapter 4

Efficient Severity Estimation with Relative Annotation

4.1 Purpose

In this chapter, I explore the application of learning to rank with relative annotation to medical image analysis. This chapter focuses on the fact that relative annotations are easier and take less time per annotation than absolute annotations. I show that learning to rank with relative annotation can reduce the cost of annotating medical images. Furthermore, to apply learning to rank to clinical practice, I propose a method to transform learning to rank with relative annotation into a regression task by calibration with a small number of absolute annotations.

As shown in Fig. 2-3 (a) in Chapter 2, the annotator attaches absolute disease levels to each image in absolute annotations. The absolute disease levels are defined based on medical findings. The disease severity can be estimated using regression and classification methods from a dataset annotated by these standards.

The task of accurately attaching discrete absolute severity levels to disease severity is difficult even for medical experts. The reason is that disease severity is inherently continuous, with gradual changes in tissues and organs, and discrete level annotations

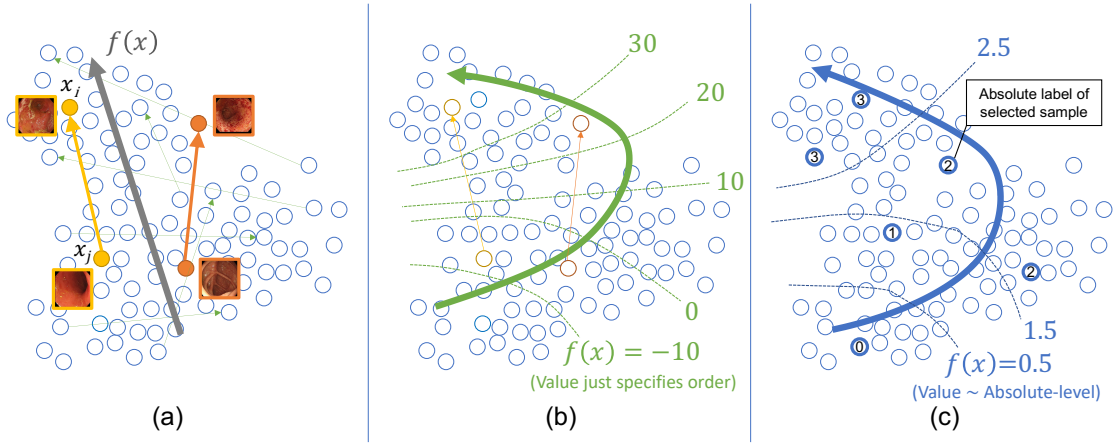


Figure 4-1: (a) Bipartite ranking problem with relative annotation and linear ranking function. (b) Nonlinear ranking function using learning to rank. (c) Nonlinear ranking function calibrated from a small number of samples with absolute annotation.

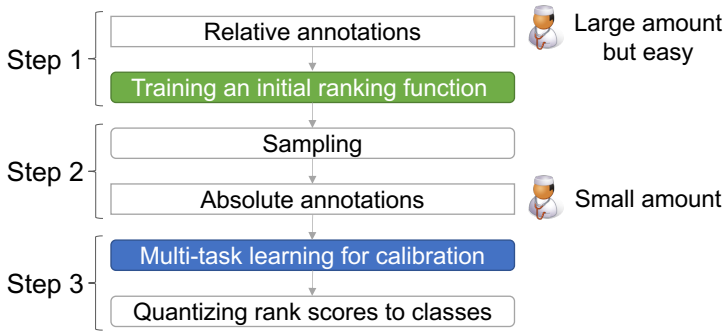


Figure 4-2: Overview of the proposed method.

such as absolute levels always result in quantization errors. For example, even if discrete annotations such as four levels of 0, 1, 2, and 3 are requested for disease severity, a severity image such as "level 1.5" could easily be found. Medical experts also consider the variability in determining disease severity due to quantization errors to be an issue in their medical practice. There are also reports of variability in levels among annotators and even within the same annotator [61].

The relative annotations shown in Fig. 2-3 (b) in Chapter 2 are simple annotations that do not require determining discrete severity levels like absolute annotations. In relative annotation, given two images (x_i, x_j), the annotator specifies the image with the higher severity. This task is much easier than compared to absolute annotations

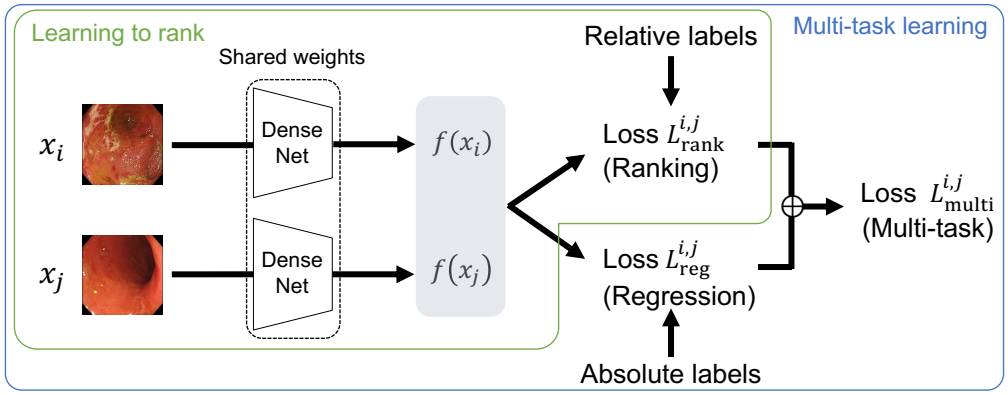


Figure 4-3: The multi-task learning in the proposed method.

and is even more accurate, especially when the difference in severity levels of the paired images is clear. Thus, the relative annotation has the potential to substantially reduce the difficulty of annotating a large number of images.

Given a dataset with the relative annotation, I can automatically estimate severity level by training a *ranking function* $f(x)$. Figure 4-1 (a) shows the so-called bipartite ranking problem idea. The basic objective of this problem is to train a function $f(x)$ that maximizes the number of sample pairs whose relative annotation is "satisfied." Specifically, suppose a pair of images x_i and x_j and a relative annotation that x_i has a higher level (i.e., ranking order) than x_j. Then, when $f(x_i) > f(x_j)$, the annotation $x_i > x_j$ is satisfied.The learned function f is expected to be relative to the original (continuous) disease level, and f enables relative severity estimation from two images.

The relationship between the original image's features and the disease's severity is generally nonlinear, so the function f should be highly nonlinear to satisfy the relative annotations as much as possible. Therefore, I use representation learning with a convolutional neural network (CNN) to obtain highly nonlinear f. Figure 4-1 (b) shows the nonlinear f by representation learning using CNN. The curve with the thick green arrow is the nonlinear f. The dotted line indicates the isolation of samples with the same rank value.

Furthermore, it should be noted that the above-ranking function still needs to

be improved for practical diagnosis in clinical practice. This is because although f can estimate the order among samples, f does not express absolute severity. In other words, if there is only one sample to be estimated, the obtained value itself has no specific meaning for diagnosis, and f cannot be used in clinical practice. For example, when realizing a ranking function f for the ulcerative colitis (UC) diagnosis by endoscopic images, the f value has no apparent relationship to a standard severity level such as the Mayo score. If we estimate UC severity from endoscopic images x_i and x_j with Mayo scores of 0 and 2, we cannot estimate Mayo scores, while we might estimate rank scores -100 and 35 that "satisfy" the severity order.

In this chapter, I propose a multi-task learning method using a large amount of (easy) relative annotations and a small amount of (costly) absolute annotations to obtain a *calibrated* ranking function $f(x)$. As shown in Fig. 4-1 (c), the ranking function f obtained by the proposed method satisfies the relative annotation $y \sim f(x)$ while satisfying for a sample x with absolute annotation y. The ranking function f can estimate *real-valued* absolute levels (e.g., Mayo scores) for *all* samples by calibration.

Figure 4-2 shows the three steps of the proposed method. In the first step, as shown in Fig. 4-1 (b), an initial ranking function is obtained by learning to rank using only relative annotations. In the second step, a small number of samples are selected based on the estimated ranking scores, and absolute annotations are attached to these samples by humans (e.g., medical experts). In the final step, as shown in Fig. 4-3, the ranking function of Step 1 is calibrated by multi-task learning using learning-to-rank and regression. The calibrated ranking function f gives real-valued disease severity levels. If the target severity level is discrete, the scores can be quantized to several levels, such as mayo 0, 1, 2, and 3. In other words, the calibrated ranking function can be regarded as a severity classifier.

I evaluated the performance of the proposed method by a classification task to estimate the UC severity level. Specifically, I used the proposed method to obtain f that estimates the Mayo score of an endoscopic image x, where the Mayo score is

represented by discrete values from 0 (normal) to 3 (severe). The f trained by the proposed method achieved higher classification performance (accuracy and F1 score) with much less annotation than the conventional multi-class classification method.

The main contributions of this chapter are summarized as follows:

1. The research I have conducted is the first attempt to use learning-to-rank with relative annotation to reduce the annotation cost for medical image datasets substantially.

2. I developed a new multi-task learning method that calibrates rank scores, which are relative severity, to absolute disease severity levels.

3. Through the UC severity classification task experiments, the proposed method was demonstrated to have equal or better classification performance than the conventional multi-class classification method , with an annotation cost of less than 1/10.

4.2 Related Work

Recently, there have been studies supporting endoscopic imaging diagnostics using machine learning. Endoscopy is used as an effective visual diagnostic test for gastrointestinal diseases. A medical expert must diagnose endoscopic images, and many studies have performed automated diagnosis using machine learning as a supplementary role to reduce the burden on medical experts. For example, as classification tasks, studies on automatic diagnosis have been reported, such as gastric cancer classification [62, 63, 64], gastric precancerous disease classification [65, 66], and colorectal cancer classification [67, 68]. As detection tasks, there have been reports of polyp detection [69, 70, 71] and early gastric cancer detection [72, 73]. Studies have also been conducted using capsule endoscopy images for automatic anomaly detection tasks [74, 75, 76]. These machine learning studies for endoscopic image analysis aim at

high diagnostic performance on classification, segmentation, and anomaly detection but not at annotation cost reduction for dataset creation, as in this chapter.

Much research on the automatic estimation of UC severity levels using deep learning has been reported. UC is a diffuse non-specific inflammatory bowel disease with repeated relapses and remissions in the colon. Accurate assessment of treatment effectiveness is critical because the type and dose of therapeutic medicines are adjusted according to the condition of patients with UC. Therefore, UC severity estimation by deep learning is expected to support medical doctors for accurate level determination. On the other hand, UC severity estimation by deep learning requires annotation on many endoscopic images for training.

There are many studies to detect UC severity by endoscopic video frames automatically [77, 78, 79, 80, 81, 82]. Takenaka et al. have studied the estimation of UC severity in biopsy images from endoscopic images [83]. Gottlieb et al. and Luo et al. proposed a method to estimate UC severity levels using CNN and a recurrent neural network (RNN) [81, 84]. The primary purpose of these methods is to estimate UC severity accurately, and the studies are not focused on reducing annotation costs.

There are also reported UC severity estimation studies that have explored reducing annotation costs. Turan et al. proposed a method using data augmentation to generate UC endoscope images using a generative adversarial network (GAN) to reduce annotation costs [85]. Schwab et al. proposed an automatic UC severity estimation method to assess treatment effectiveness [86]. Their method employs weakly supervised learning, as it would be too costly to annotate every captured image fully. On the other hand, unlike our method, the methods of Turan et al. and Schwab et al. use only absolute annotation and do not solve the issue of label variability and annotation difficulties, as described in Chapter 2.

4.3 Learning to Rank with Calibration

4.3.1 Two Types of Ground Truth Labels

This chapter treats two types of ground truth labels for learning: absolute labels (AL) and relative labels (RL). AL is the ground truth attached to a single image by absolute annotation, and RL is the ground truth attached to two image pairs by relative annotation. As shown in Fig. 4-2, the proposed method first attaches RLs to a large number of image pairs and second attaches ALs to a small number of images.

4.3.2 Overview

The proposed method consists of three steps. In Step 1, I first train the initial ranking function by using learning to rank with RLs. In Step 2, I select small samples from the training data to annotate them regarding ALs. The samples are selected using the ranking function trained using the RLs. In Step 3, I finally carried out multi-task learning with RLs and additionally and adaptively prepared ALs, for calibrating the ranking function to more meaningful disease-severity levels. An overview of the proposed method is shown in Fig. 4-2.

Before providing further details of the above steps, two important aspects should be clarified. First, the rank score by the ranking function in Step 1 is not a disease-severity level, and thus the calibration of Step 3 is necessary. Second, an AL is not given in advance but given after the ranking function is obtained using RLs. This provides a more appropriate choice of samples where ALs should be attached, resulting in more accurate severity-level estimation with less AL annotation cost.

4.3.3 Details

Learning to rank

In Step 1, I train the initial ranking function using learning to rank. The ranking function $f(x)$ is trained with a CNN for the representation (i.e., feature extraction) that is suitable for the ranking. A CNN is composed of multiple convolutional layers, a single fully connected layer, and a single output node to give a single scalar value $f(x)$. This can be considered a powerful extension of the classical RankNet [14], where a linear ranking function is trained using a very shallow neural network.

The CNN is trained using sample pairs with RLs. For training, I input two images to two CNNs with shared weights and then minimize the loss for the pair. Specifically, the CNN is trained with the loss function $L_{\text{rank}} = \sum_{(i,j)\in\mathcal{P}} L^{i,j}_{\text{rank}}$, where $\mathcal{P}$ is the set of sample pairs. The function $L^{i,j}_{\text{rank}}$ is defined as a cross-entropy,

$$L^{i,j}_{\text{rank}} = -\bar{P}_{i,j} \log P_{i,j} - (1 - \bar{P}_{i,j}) \log(1 - P_{i,j}), \tag{4.1}$$

where $P_{ij} = \text{sigmoid}(f(x_i) - f(x_j))$ and $\bar{P}_{ij}$ is an RL of the image pair (x_i, x_j).

Sampling for absolute annotation

In Step 2, small samples are selected from the training data and attached ALs. As noted above, the proposed method assumes that the samples to which ALs are attached are selected *after* $f(x)$ is estimated. This is more reasonable than, for example, a random selection because I can select samples that are expected to be more necessary for the calibration step by using the clues from $f(x)$.

Figure 4-4 shows the selection method of samples for AL annotation. The line circles indicate the training data, and the red circles indicate the training data to be selected for attaching absolute labels. To select a small number of samples, I first obtain the rank score of the training samples using $f(x)$ and represent the rank score of the training data as a point on a number line. Next, I select M ($\ll N$) samples at

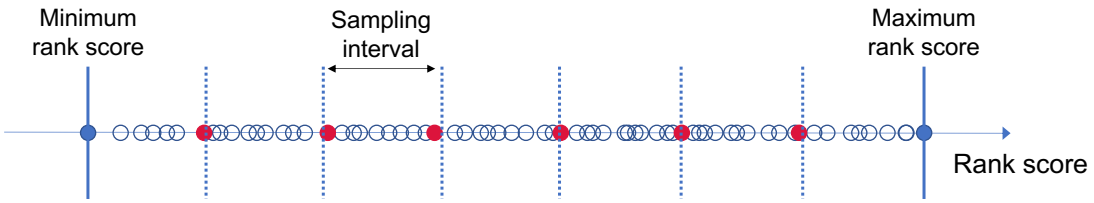


Figure 4-4: Selection of samples for AL annotation.

equal intervals on the number line within the maximum and minimum rank scores. Finally, absolute annotations attach ALs to the selected M samples.

Multi-task learning

In the final Step 3, the proposed method calibrates the ranking function to give the absolute severity score. This calibration process can be seen as a fine-tuning process of $f(x)$ so that the output of $f(x)$ becomes closer to the AL of x. At the same time, I need to be careful that the fine-tuning process does not destroy the sample ranks learned in $f(x)$. These two requirements result in multi-task learning to fine-tune $f(x)$.

As shown in Fig. 4-3, the multi-task learning combines regression to make $y \sim f(x)$ for the sample x with AL and learning to rank for the pairs (x_i, x_j) with RL. The loss function of learning to rank is the cross entropy of Eq.(4.1). The loss function for regression, $L_{\text{reg}} = \sum_{(i,j)\in\mathcal{P}} L_{\text{reg}}^{i,j}$ is defined by adding the mean squared error (MSE) loss function for each sample pair,

$$L_{\text{reg}}^{i,j} = (f(x_i) - y_i)^2 + (f(x_j) - y_j)^2, \tag{4.2}$$

where y_i and y_j are the ALs attached by the absolute annotation of x_i and x_j, respectively. Furthermore, the multi-task loss function L_{multi} is defined as the sum of the loss functions of learning to rank and regression,

$$L_{\text{multi}} = L_{\text{rank}} + \lambda L_{\text{reg}}, \tag{4.3}$$

where λ is a hyper-parameter to balance the losses. The trained multi-task $f(x)$ is expected to be a ranking function corrected for the region of severity levels in the feature space optimized by representation learning, as shown in Fig. 4-1 (c). Note that, in Step 3, I only use M samples with ALs. Therefore, L_{rank} in Eq.(4.3) is minimized with the RLs of $M(M-1)/2$ pairs.

Finally, the calibrated rank score $f(x)$ is quantized into the nearest discrete disease severity level as the classification result of x. For example, with the severity levels $\in \{0, 1, 2, 3\}$, the level becomes 3 for x whose $f(x) = 2.7$.

4.4 Experiments and Results

4.4.1 Dataset

I used 10,265 colonoscopy images of UC from 388 patients at Kyoto Second Red Cross Hospital as the dataset. These images were taken from multiple patients (including healthy participants). The images have different sizes and therefore were resized to 224×224 pixels.

Figure 4-5 shows several examples of each of the four levels of Mayo, which is the standard disease severity score for UC. According to Schroeder et al. [46], Mayo 0 is normal or endoscopic remission. Mayo 1 is a mild level showing erythema (i.e., abnormal redness), a decreased vascular pattern, and mild friability. Mayo 2 is a moderate level showing marked erythema, an absent vascular pattern, friability, and erosions. Mayo 3 is a severe level with spontaneous bleeding and ulceration. AL is a disease severity level as generally defined in medicine. Therefore, since this chapter uses the UC endoscopic image dataset, AL corresponds to one of the four Mayo scores.

Although our method does not require a dataset with full ALs, medical experts attached ALs to all samples for a quantitative performance evaluation. Specifically, a four-level Mayo score is carefully attached to each colonoscopy image by multiple medical experts. The dataset contains 6,678, 1,995, 1,395, and 197 samples for

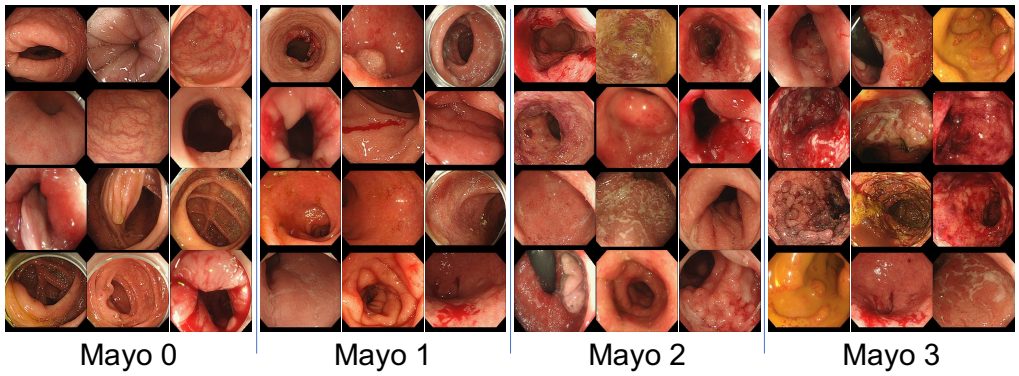


Figure 4-5: Colonoscopy image examples with different UC levels.

Mayo 0, 1, 2, and 3, respectively. Note that it is common to have such a heavily imbalanced dataset for colonoscopy, as well as other medical image diagnosis tasks.

In the following experiments, five-fold cross-validation was performed. The colonoscopy images were divided into 60%, 20%, and 20% for patient-disjoint training, validation, and test sets, respectively. Note that I divide all images into training, validation, and test sets to have the same severity proportion for keeping a fair and practical evaluation scenario.

I carefully control several conditions for a fair comparison with a conventional method (detailed later). First, I used the same number of annotations for the conventional and proposed methods. More precisely, the conventional method uses 8,212 images-(80% of the entire data) with AL for training, and the proposed method uses $8{,}212-M$ pairs with RL at Step 1, and M images with AL at Step 3. This condition is a handicap for the proposed method because AL has more information than RL in the classification task. Nevertheless, I adopted this condition so that the conventional method would not be disadvantaged.

Second, I allow over/under sampling for class imbalance removal to the conventional method but not the proposed method. This is because the conventional method has ALs for all samples; thus, such sampling is possible, whereas the proposed method does not. This condition will be another large handicap for the proposed method.

Table 4.1: Comparison of ground-truth labels and annotation time for each method.

Method	RL	AL	Time (sec)*
Conventional	-	8,212	164,240
Proposed	8,212$-M^{**}$	M^{**}	8,212+19M^{**}

* RL: 1 (sec/pair), AL: 20 (sec/image)
** M = 50, 100, 200, 300, 400

4.4.2 Effects of Annotation Cost Reduction

Table 4.1 shows the number of ground-truth labels (RL and AL) and the annotation time for each method. In our interview with endoscopists, AL labeling takes 20 seconds per image, and RL labeling only takes (less than but roughly) one second. For the case $M = 400$, this indicates that the proposed method requires just 10% of the annotation time of the conventional method.

4.4.3 Implementation

The implementation environment is shown as follows. I used an Intel(R) Core(TM) i9-10980XE 3.00 GHz as the CPU and two NVIDIA TITAN RTX 24 GB as GPUs for training. I wrote the code in Python 3.6 and used Tensorflow 1.13.1 and Keras 2.2.4 as the deep learning library. The CUDA version was 10.0. I used Adam as the optimizer to train the weight parameters. The learning rate was set to 5×10^{-6}. The learning was terminated by the early stopping rule (no decrease in validation loss for 20 epochs). For λ in Eq.(4.3), I examined the range of 0.001 to 1 and had the highest F1-score at $\lambda = 0.01$ for the validation set.

I used DenseNet-169 [87] as the CNN. DenseNet has been widely used in various medical-image classification and analysis tasks due to its state-of-the-art performance (e.g., [88, 89]).

4.4.4 Evaluation Metrics

The proposed method is evaluated in four-Mayo class classification performance by accuracy, recall, precision, and F1-score. Recall that the class is determined by quantizing the rank score into its neighboring level, e.g., $2.7 \rightarrow 3$. I leave the test samples imbalanced to mimic realistic medical situations. To avoid the under/over-estimation risk of the accuracy values in the imbalanced situation, F1-score is also employed.

4.4.5 Comparison Method

The performance of the proposed method was compared with the conventional CNN-based multi-class classification method. DenseNet-169, trained by the standard categorical cross entropy, is used for this comparative method. As for the training data, all 8,212 training samples are used with their ALs. This means that it uses all of the ALs.

4.4.6 Classification Performance Results

Table 4.2 shows the classification performance of the proposed method at $M = 400$. The proposed method achieves a higher F1-score than the conventional method. This result shows that the proposed method achieves even higher classification performance than the conventional method, although the proposed method only needs the 1/10 annotation cost of the conventional method.

I evaluated the performance of the proposed method for severity classification with various numbers of ALs ($M = 50$, 100, 200, 300, and 400). Figure 4-6 shows the performance evaluation results with the proposed method using different numbers of ALs. *Acc* and *F1* represent accuracy and F1-score, respectively. The F1-score increases as the number of ALs increases. From $M = 300$, the F1-score of the proposed method is higher than that of the conventional method. The accuracy of the proposed method is higher than that of the conventional method for $M = 200$ and over.

Table 4.2: Classification performance evaluation of the conventional and proposed methods.

Method	Class	Precision	Recall	F1-score
Conventional	Mayo 0	0.901	0.747	0.817
	Mayo 1	0.394	0.675	0.498
	Mayo 2	0.764	0.443	0.561
	Mayo 3	0.252	0.635	0.360
	Overall	0.578	0.625	0.559
Proposed	Mayo 0	0.864	0.840	0.852
	Mayo 1	0.416	0.463	0.438
	Mayo 2	0.643	0.606	0.624
	Mayo 3	0.368	0.432	0.397
	Overall	0.572	0.585	**0.578**

Therefore, the proposed method achieved higher performance than the conventional method when the number of ALs is more than $M = 300$.

4.4.7 Ablation Study

I examined the effect of calibration on rank scores by multi-task learning with the proposed method. Specifically, I verified the effect by comparing the classification performance between calibrated and uncalibrated cases. To determine the classification result for the uncalibrated case, I defined the range of the rank score for each Mayo score by logistic regression using $M = 400$ ALs, which were used for multi-task learning, for a fair comparison.

Figure 4-7 shows box-plots for each correct Mayo score of test samples by (a) the uncalibrated case and (b) the proposed method. The horizontal and vertical axes correspond to the correct Mayo score attached by the annotators and the rank scores, respectively. The rank scores obtained with the proposed method are located nearer to the Mayo score range of 0 to 3 than those with the uncalibrated case. Therefore, these results indicate that the rank scores are calibrated with the ALs as the anchor by using regression as the anchor task in multi-task learning.

Table 4.3 shows the results of the performance evaluation for the uncalibrated

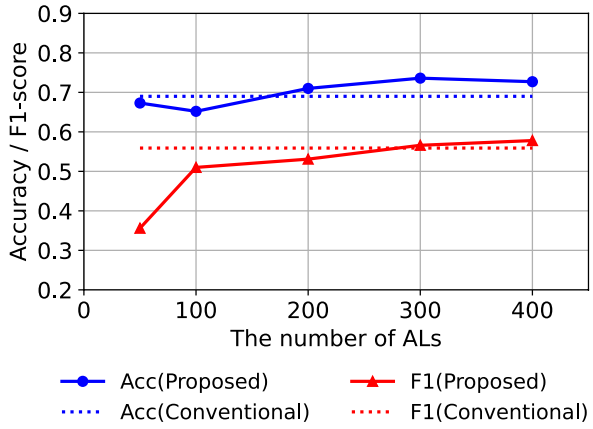


Figure 4-6: Classification performance evaluation of the proposed method using different numbers of ALs. In the conventional method, 8,212 ALs are used.

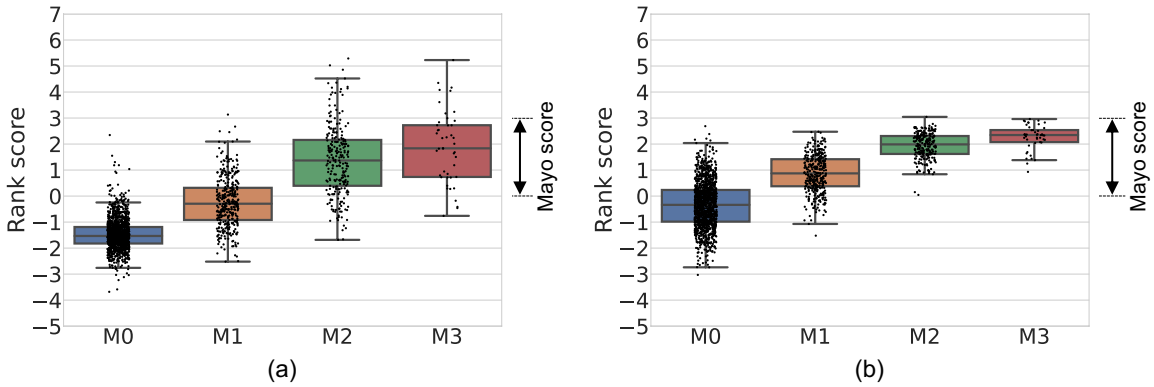


Figure 4-7: Box-plots of test set's rank score obtained with (a) the uncalibrated case and (b) the proposed method. 'M0' denotes Mayo 0.

case. The overall precision, recall, and F1-score of the uncalibrated case were lower than those of the proposed method. Comparing the F1-score for each class, Mayo 3 was particularly low with the uncalibrated case, indicating an imbalance in the classification performance for each class.

Figure 4-8 shows the confusion matrices of the classification results using the uncalibrated case and the proposed method. Compared with the proposed method, the uncalibrated case had a higher rate of incorrectly predicting Mayo 2 and Mayo 3 as Mayo 1 and could not accurately classify images with high severity. These results

Table 4.3: Classification performance evaluation of the uncalibrated case.

Method	Class	Precision	Recall	F1-score
Uncalibrated	Mayo 0	0.892	0.873	0.882
	Mayo 1	0.439	0.625	0.516
	Mayo 2	0.728	0.422	0.535
	Mayo 3	0.229	0.086	0.124
	Overall	0.570	0.502	0.514

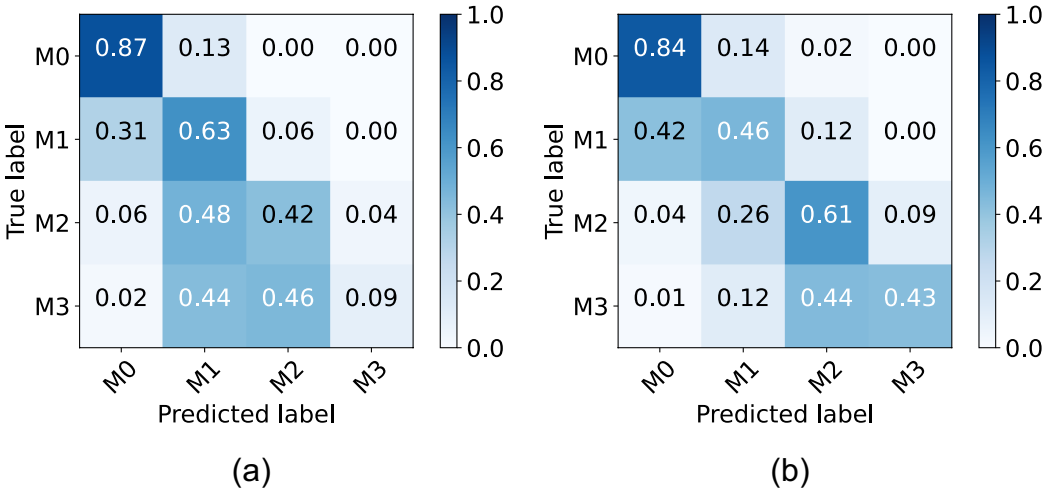


Figure 4-8: Confusion matrices of (a) the uncalibrated case and (b) the proposed method.

indicate that the calibration effect improves the performance of classifying images with high severity and that the proposed method has higher performance than the uncalibrated case.

4.5 Limitation

There are three main limitations of the proposed method. First, the number of pairs increases quadratically with the number of samples in the experiments of Step 1 in Fig. 4-2, an essential issue for relative annotation. The number of annotatable pairs for N samples increases to $N(N-1)/2$. Therefore, it is practically impossible to annotate all pairs when the number of samples is large. Although N pairs were

selected randomly from N samples, it is necessary to consider a more effective method for creating pairs.

Second, the relative annotation of Step 1 in Fig. 4-2 cannot deal with class imbalance. Relative annotation compares the severity of two samples and thus cannot determine the severity proportion in the entire dataset. Therefore, it is impossible to balance by over-sampling or under-sampling based on the severity proportion or adjust the loss function by the class weights to deal with class imbalance when learning. Since the proposed method can not address the class imbalance in relative annotation, a ranking function with high accuracy can not be obtained. Therefore, it is necessary to consider pair-creation methods that address the class imbalance.

Third, in the experiments of Step 3 in Fig. 4-2, it takes a long time to train a model because all combination pairs were created from the sampled images of Step 2. For example, the number of pairs that can be created was about 80,000 pairs for $M = 400$. When using all combination pairs in training, the proposed method is limited in the number of samplings in Step 3 in terms of computational cost. Therefore, it is necessary to consider a method to create pairs with high-learning effectiveness, even in multi-task learning in Step 3.

4.6 Conclusion

I proposed a multi-task learning method that combines learning to rank with regression for automatically estimating UC severity levels (Mayo scores). The proposed method has a strong advantage in substantially reducing annotation costs by using relative annotations instead of costly absolute annotations. Our experimental result shows that the proposed method achieved even higher classification performance (accuracy and F1-score) than the conventional classification method while requiring just 1/10 annotation costs.

Future work will involve the proposal of a new continuous-valued UC severity level.

Currently, I discretize the regression result into four levels to follow the traditional Mayo-based evaluation. However, the regression result can show an intermediate score, such as Mayo 1.75 by itself. Our continuous severity score can be an accurate and precise alternative to Mayo through discussion with the medical expert committee.

Chapter 5

Deep Bayesian Active-Learning-to-Rank for Medical Images

5.1 Purpose

In this chapter, I address the optimal pair selection issue for relative annotation. The previous chapters have demonstrated that learning to rank by relative annotation can substantially reduce the annotation cost in medical image analysis and achieve higher performance than conventional multi-class classification methods. This chapter focuses on the issue that the number of pairs that can be attached to a label is too large for relative annotation, and thus pairs must be selected that are appropriate for the high learning effect.

Relative annotation is expected to be a promising annotation for image-based severity estimation with lower annotation cost than absolute annotation. Absolute annotation, which attaches a discrete label to a single image, is considered difficult even by medical experts because there are often images of intermediate severity between each discrete level. On the other hand, relative annotation, which attaches

labels by comparing the severity of two images, is easier and has less variability.

However, as described in Section 4.5, the relative annotation has the issue that the number of pairs increases quadratically with the number of images. Image analysis using deep learning requires a large amount of training data. The number of pairs becomes too large when the number of collected images is significant, and thus it is difficult to annotate all pairs in practice. Therefore, it is important to select pairs with high learning effectiveness for relative annotation.

The relative annotation also has the issue that it cannot deal with the class imbalance in the dataset when creating pairs for annotation. A naive sample selection strategy when pair creation is a random selection, and the pair creation on class-imbalanced datasets increases the number of pairs consisting of samples in the majority class. In other words, the random selection easily overlooks minority but important samples, which often appear in medical applications. It is necessary to select samples to mitigate the class imbalance of the training dataset when creating pairs for relative annotations.

In this chapter, I propose a deep Bayesian active-learning-to-rank that fully maximizes the efficiency of relative annotation for efficient image-based severity estimation. The technical novelty of the proposed method is to introduce an *active learning* technique into the learning-to-rank framework for selecting a less number of effective sample pairs. Active learning has been studied [90] and generally takes the following steps. Firstly, a neural network is trained with a small amount of annotated samples. The trained network then suggests other samples to be annotated. These steps are repeated to have enough amount of annotated training samples.

For suggesting important samples to be annotated, I employ an *uncertainty* of the samples. If I find two samples with high uncertainty about their rank score, a new relative annotation between them will be useful to boost the ranking performance. For this purpose, I employ Bayesian CNN [91] to realize a ranking function because it can give an uncertainty of each sample. In summary, our deep Bayesian active-learning-

to-rank uses Bayesian CNN to rank with representation and uncertainty-based active learning.

As an experimental validation of the efficiency of the proposed method, I perform UC severity estimation using approximately 10,000 endoscopic images collected from the Kyoto Second Red Cross Hospital. Quantitative and qualitative evaluations clearly show the efficiency of the proposed framework. Especially, I reveal that the proposed method automatically selects minority but important samples.

The main contributions of this chapter are summarized as follows:

- I propose a deep Bayesian active-learning-to-rank that can rank the image samples according to their severity level. The proposed method can automatically select important image pairs for relative annotation, resulting in learning with a less number of relative annotations.

- Through UC severity estimation experiments, I demonstrated that the proposed method could suppress the number of relative annotations while keeping ranking performance.

- I also experimentally confirmed that the proposed method shows robustness to class imbalance by its ability to select minor but important samples.

5.2 Related Work

Deep learning requires a large amount of training data, but annotating all the unlabeled data can sometimes be difficult. In this situation, active learning that selects and annotates unlabeled data with a high learning effect is suitable. Active learning is mainly diversity-based sampling and uncertainty-based sampling. Many diversity-based sampling methods have been proposed for active learning in image analysis [92, 93, 94, 95]. These methods select high-diversity samples that are not similar from different clusters or that are distant from each other in feature space.

However, these methods are used for absolute annotation and do not assume image pair diversity for relative annotation.

Many uncertainty-based sampling methods have been proposed for active learning in image analysis [91, 96, 97, 98, 99, 100]. Gal et al. proposed active learning with uncertainty based on Bayesian CNN [91]. The model uncertainty they proposed has been adopted in many active learning methods. Nair et al. [99] proposed a method using active learning for medical image analysis with uncertainty obtained from a Bayesian CNN. This method deals with an orthodox classification task with absolute annotation and thus does not assume any relative annotation and learning-to-rank.

Many active learning methods combine uncertainty-based and diversity-based sampling [101, 102, 103]. Uncertainty-based sampling selects under-trained samples but selects more similar images. Diversity-based sampling selects samples that are not similar but selects more samples that are easier to estimate with the model. Therefore, many methods have been proposed to overcome the disadvantages and utilize the advantages of each sampling method.

In the natural language process (NLP) field, Wang et al. [104] recently introduced deep Bayesian active learning to learning-to-rank. Although their method sounds similar to ours, its aim, structure, and application are totally different from ours. Specifically, it aims to rank sentences (called answers) to fixed sentences (called a query) according to their relevance. For this aim, its network always takes two inputs (like $f(x_i, x_j)$), whereas ours takes a single input (like $f(x_i)$). From these differences, it is impossible to use it for the active annotation task and even to compare ours with it.

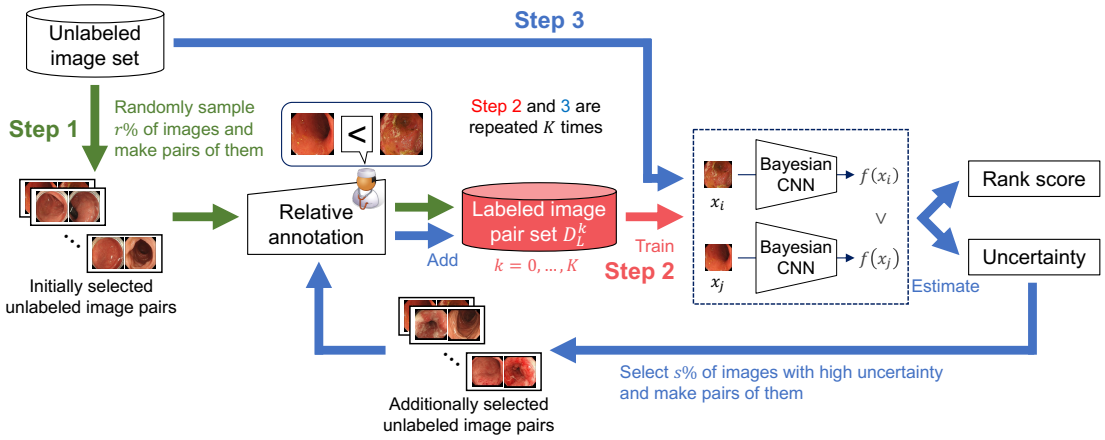


Figure 5-1: Overall structure of proposed deep Bayesian active-learning-to-rank for severity estimation of medical image data.

5.3 Deep Bayesian Active-Learning-to-Rank

5.3.1 Overview

Figure 5-1 shows an overall structure of the proposed deep Bayesian active-learning-to-rank. The proposed method is organized in an active learning framework using Bayesian learning and learning-to-rank for severity estimation of medical image data; experts progressively add relative annotations to image pairs selected based on the uncertainty provided by a Bayesian CNN while training the Bayesian CNN based on the learning-to-rank algorithm.

The deep Bayesian active-learning-to-rank consists of three steps. In Step 1, for initial training, a small number of image pairs are randomly sampled from the unlabeled image set and then annotated by medical experts. In Step 2, a Bayesian CNN is trained with the labeled image pairs for estimating rank scores and the uncertainties of individual training samples. In Step 3, images to be additionally labeled are selected based on estimated uncertainties, having the medical experts annotate the additionally selected images via relative annotation. Steps 2 to 3 are repeated K times to train the Bayesian CNN while progressively increasing the number of labeled image pairs.

5.3.2 Details

More details of each step are explained as follows.

Step 1: Preparing a small number of image pairs for initial training. A set with a small number of annotated image pairs $\mathcal{D}_{\mathrm{L}}^0$ is prepared for initial training as follows. First, given an unlabeled image set comprising N unlabeled images, I randomly sample $r\%$ of them (i.e., $R = rN/100$ image samples). Then, to form a set of R pairs (instead of all possible $R(R-1)/2$ pairs), one of the $R-1$ samples is randomly paired for each of the R samples[1]. Relative labels are attached to these image pairs by medical experts. For an image pair (x_i, x_j), the relative label $C_{i,j}$ is defined by the three values (1, 0.5, 0) as described in Chapter 2. After this step, I have the annotated image pair set $\mathcal{D}_{\mathrm{L}}^0 = \{(x_i, x_j, C_{i,j})\}$ and $|\mathcal{D}_{\mathrm{L}}^0| = R$.

Step 2: Training a Bayesian CNN. A Bayesian CNN is trained as a ranking function with the labeled image pair set $\mathcal{D}_{\mathrm{L}}^0$. The CNN outputs a rank score that predicts the severity of the input image and the prediction's uncertainty. In training, I employ a probabilistic ranking cost function [14] to conduct learning-to-rank with a neural network and Monte Carlo (MC) dropout [105, 91] for approximate Bayesian inference.

Let $f(\cdot)$ be a ranking function by a CNN with L weighted layers. Given an image x, the CNN outputs a scalar value $f(x)$ as the rank score of x. I denote by $\mathbf{W}_l$ the l-th weight tensor of the CNN. For a minibatch $\mathcal{M}$ sampled from $\mathcal{D}_{\mathrm{L}}^0$, the Bayesian CNN is trained while conducting dropout with the loss function $\mathcal{L}_{\mathcal{M}}$ defined as follows:

$$\mathcal{L}_{\mathcal{M}} = -\sum_{(i,j)\in\mathcal{I}_{\mathcal{M}}} \{C_{i,j}\log P_{i,j} + (1-C_{i,j})\log(1-P_{i,j})\} + \lambda\sum_{l=1}^{L}\|\mathbf{W}_l\|_F^2, \qquad (5.1)$$

where $\mathcal{I}_{\mathcal{M}}$ is a set of index pairs of the elements in $\mathcal{M}$, $P_{i,j} = \mathrm{sigmoid}(f(x_i) - f(x_j))$, λ is a constant value for weight decay, and $\|\cdot\|_F$ represents a Frobenius norm. In

[1]The strategy of making pairs is arbitrary. Here, I want to annotate all of the R samples at least one time while avoiding $O(R^2)$ annotations and thus take this strategy.

Eq. (5.1), the first term is a probabilistic ranking cost function [14], which allows the CNN to train rank scores, and the second term is a weight regularization term that can be derived from the Kullback–Leibler divergence between the approximate posterior and the posterior of the CNN weights [105]. The CNN is trained by minimizing the loss function $\mathcal{L}_{\mathcal{M}}$ for every minibatch while conducting dropout; a binary random variable that takes one with a probability of p_{dropout} is sampled for every unit in the CNN at each forward calculation, and the output of the unit is set to zero if the corresponding binary variable takes zero.

The rank score for an unlabeled image x^* is predicted by averaging over the output of the trained Bayesian CNN with MC dropout as $y^* = \frac{1}{T}\sum_{t=1}^{T} f(x^*; \omega_t)$, where T is the number of MC dropout trials, ω_t is the t-th realization of a set of the CNN weights obtained by MC dropout, and $f(\cdot; \omega_t)$ is the output of $f(\cdot)$ given a weight set ω_t.

The uncertainty of the prediction is defined as the variance of the posterior distribution of y^*. This uncertainty is used to select images annotated in the next step, which plays an important role in achieving active learning. The variance of the posterior distribution, $\mathrm{Var}_{q(y^*|x^*)}[y^*]$, is approximately calculated using MC dropout as follows:

$$\begin{aligned}\mathrm{Var}_{q(y^*|x^*)}[y^*] &= \mathbb{E}_{q(y^*|x^*)}[(y^*)^2] - \left(\mathbb{E}_{q(y^*|x^*)}[y^*]\right)^2 \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\left(f(x^*; \omega_t)\right)^2 - \left(\frac{1}{T}\sum_{t=1}^{T} f(x^*; \omega_t)\right)^2 + \text{const.}, \end{aligned} \tag{5.2}$$

where $q(y^*|x^*)$ is the posterior distribution estimated by the model. The constant term can be ignored because the absolute value of uncertainty is not required in the following step.

Step 3: Uncertainty-based sample selection. A new set of annotated image pairs is provided based on the estimated uncertainty and relative annotation. I es-

timate the rank scores and the related uncertainties for the unlabeled images, using the trained Bayesian CNN, and select $s\%$ images with high uncertainty. Image pairs are made by pairing the selected images in the same manner as Step 1, and medical experts attach relative annotations to the image pairs. The newly annotated image pairs are added to the current annotated image pair set $\mathcal{D}_{\mathrm{L}}^{0}$. The Bayesian CNN is retrained with the updated $\mathcal{D}_{\mathrm{L}}^{1}$. Steps 2 and 3 are repeated K times while increasing the size of the annotated set $\mathcal{D}_{\mathrm{L}}^{k}$ $(k = 0, \ldots, K)$.

5.4 Experiments and Results

To evaluate the efficiency of our active learning-to-rank, I conducted experiments on a task for estimating rank scores of ulcerative colitis (UC) severity. In the experiments, I quantitatively compared our method with baseline methods in terms of the accuracy of relative label estimation, that is, the correctness of identifying the higher severity image in a given pair of two endoscopic images. I also evaluated the relationship between the human-annotated absolute labels and the estimated rank scores.

In addition, I analyzed the reason why our method successfully improved the performance of relative label estimation. Especially, I analyze the relationship between the uncertainty and class prior and show that our uncertainty-based sample selection could mitigate the class imbalance problem and then finally improve the performance.

5.4.1 Dataset

In this chapter, experiments were performed using the UC endoscopic image dataset collected at the Kyoto Second Red Cross Hospital, which was used in Chapter 4. The dataset consists of $10,265$ endoscopic images from 388 patients. It should be noted that the dataset has imbalanced class priors, which is a typical condition in medical image analysis. Specifically, it contains 6,678, 1,995, 1,395, and 197 samples for Mayo 0, 1, 2, and 3, respectively.

To evaluate relative label-based methods, I made a set of pairs of two images and gave relative labels for each pair based on the Mayo labels. For N training samples, the number of possible pairs is too large ($O(N^2)$) to train the network with various settings; we, therefore, made a limited number of pairs for training data by random sampling. Specifically, I used N pairs (instead of $O(N^2)$ pairs) by selecting one of the $N-1$ samples for each of the N samples. (In Section 5.3.2, R samples of the initial set are also selected from these pairs.) I consider this setting reasonable since it used all the original samples, and I could conduct the experiments in a realistic running time. Also, note that this setting is typical for evaluating learning-to-rank [106, 107, 108].

Five-fold cross-validation was conducted for all comparative methods. The data were divided into training (60%), validation (20%), and test (20%) sets by patient-based sampling (that is, each data did not contain an image from the same patient). Note that the above pair image preparation was done after dividing the training, validation, and test sets for a fair evaluation. It indicates that each set did not contain the same image. As the performance metrics, I used the accuracy of estimated relative labels, which is defined as the ratio of the number of correctly estimated relative labels over all the pairs.

5.4.2 Implementation

I used DenseNet-169 [87] as the backbone of the Bayesian CNN. The model was trained with dropout ($p_{\text{dropout}} = 0.2$) and weight decay ($\lambda = 1{\times}10^{-4}$) in the convolutional and fully connected layers. I used Adam as the optimization algorithm and set the initial learning rate to $1{\times}10^{-5}$. All image data were resized to 224×224 pixels and normalized between 0 and 255.

In all experiments, our method incrementally increased the training data during $K = 6$ iterations by selecting effective samples ($s = 5\%$ of training data) and adding them to the initial training data ($r = 20\%$ of all training images). In total (after six

iterations), the ratio of the labeled data used for training was 50% ($r+sK = 20+5\times6$). The number of estimations for uncertainty estimation was set to $T = 30$.

5.4.3 Comparison Methods

To demonstrate the effectiveness of our method, I compared our method with two baseline methods and one ablation study: 1) Baseline, which trained by randomly sampled ($r+sK =$) 50% pairs of training data, which indicates that the same number of pairs were used in the proposed method; 2) Baseline (all data), which uses all N training pairs (i.e., 100%), which indicates the number of training data was as twice as that of the proposed method; 3) Proposed w/o uncertainty-based sampling (UBS), which also incrementally increased the training data during K iterations but the additional training data was selected by random sampling (without using uncertainty-based sampling); For a fair comparison, I used the same backbone (DenseNet-169 based Bayesian CNN) for all the methods. Given an input image, all methods used the mean of rank scores of $T = 30$ times estimation as the rank score.

5.4.4 Evaluation of Relative Severity Estimation

As the accuracy evaluation, I measured the correctness of the estimated rank scores in a relative manner. Specifically, given a pair of images (x_i, x_j) where x_i shows a higher severity, the estimation is counted as "correct" if $f(x_i) > f(x_j)$.

For the test data, I prepared two types of pairs.

"Overall": The pairs were randomly made among all Mayo. Specifically, I selected images in test images so that the number of samples in each Mayo score was the same and then randomly made the pairs among the selected images. Using this test dataset, I evaluated the overall performances of the comparative methods. However, this test data may contain many easy pairs, that is, a pair of Mayo 0 (a normal image) and Mayo 3 (a high severity image), whose image features are very different, as shown in Fig. 4-5 of Chapter 4. Thus it is easy to identify the relative label.

Table 5.1: Quantitative performance evaluation in terms of accuracy of estimated relative labels. '*' denotes a statistically significant difference between the proposed method and each comparison method ($p < 0.05$ in McNemar's test).

Method	Labeling ratio	Overall	Neighboring			
			0–1	1–2	2–3	Mean
Baseline	50%	0.861*	0.827	0.837	0.628	0.763*
Baseline (all data)	100%	0.875	**0.855**	0.870	0.635	0.785*
Proposed w/o UBS	50%	0.856*	0.818	0.842	0.634	0.763*
Proposed	50%	**0.880**	0.787	**0.871**	**0.736**	**0.797**

"Neighboring": I prepared the neighboring pairs that contain the images of neighboring level Mayo, such as Mayo 0–1, Mayo 1–2, and Mayo 2–3 pairs. It is important for clinical applications to compare the severity of difficult cases. This estimation is more difficult since the severity gradually changes in neighboring Mayo, and these image features are similar. Using this test dataset, I evaluated the performance of methods in difficult cases.

Table 5.1 shows the mean of the accuracy of estimated rank scores for each method in five-fold cross-validation. A labeling ratio denotes the ratio of the number of labeled images that were used for training, and '*' indicates that there were significant differences at $p < 0.05$ by multiple statistical comparisons using McNemar's tests.

In the results of "Overall," the accuracy of Baseline and Proposed w/o UBS were comparable because these methods used the same number of training pairs. Baseline (all data) improved the accuracy compared to these two methods by using the larger size (twice) of training data. Surprisingly, our method was better than Baseline (all data), although our method's training data is half that of Baseline (all data) with the same settings, that is, the network structure and the loss. As analyzed in the later Section 5.4.5, this difference comes from class imbalance. Since this dataset has a severe class imbalance, that is, the samples in Mayo 0 were 33 times of those in Mayo 3, it was difficult to learn the image features of highly severe images. In such severe imbalance cases, even if the number of training data decreases, appropriate sampling could mitigate the class imbalance and improve the accuracy.

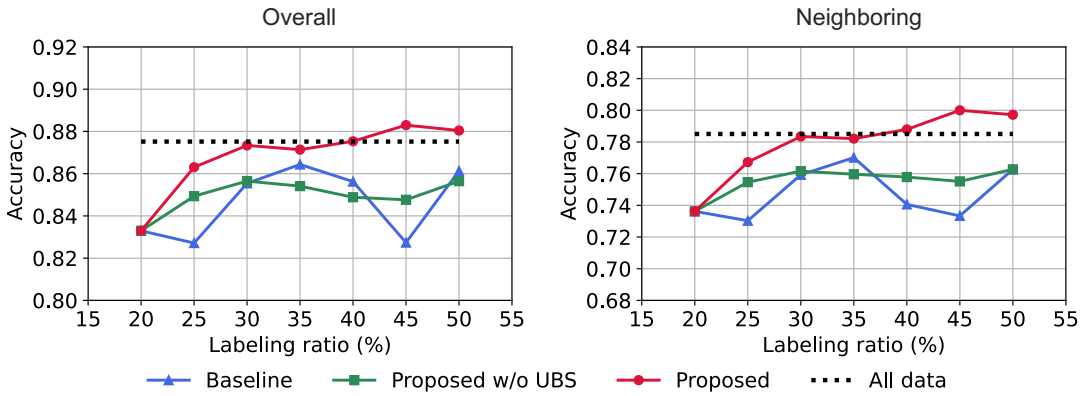


Figure 5-2: Accuracy of estimated relative labels of Baseline (blue), Proposed w/o UBS (green), and Proposed (red) at each labeling ratio. The black dotted line indicates the results of Baseline (all data).

In the results of "Neighboring" (difficult cases), the proposed method was also better than all the comparative methods in the mean accuracy of the neighboring pairs. In particular, the proposed method improved the accuracy in the case "Mayo 2–3" over 10%. Since the accuracy of image comparisons at high severity is important for evaluating treatment effects, the proposed method is considered superior to the other methods in clinical practice. In the case "Mayo 0–1", the accuracy of the proposed method was lower than that of the comparison methods because the number of labeled samples for Mayo 0 and 1 was reduced due to the mitigation of class imbalance.

Figure 5-2 shows the changes in the accuracy at each iteration in both test datasets "Overall" and "Neighboring." The horizontal axis indicates the labeling ratio, the vertical axis indicates the mean accuracy in cross-validation, and the black dot line indicates the results of Baseline (all data), which used the 100% of the training data. This result shows the effectiveness of our uncertainty-based active learning. Our method (red) increased the accuracy with the number of training data, and the improvement was larger than the other methods. In contrast, for Baseline (blue) and Proposed w/o UBS (green), it was not always true to increase the accuracy by increasing the training data. As a result, the improvements from the initial training data were limited for them.

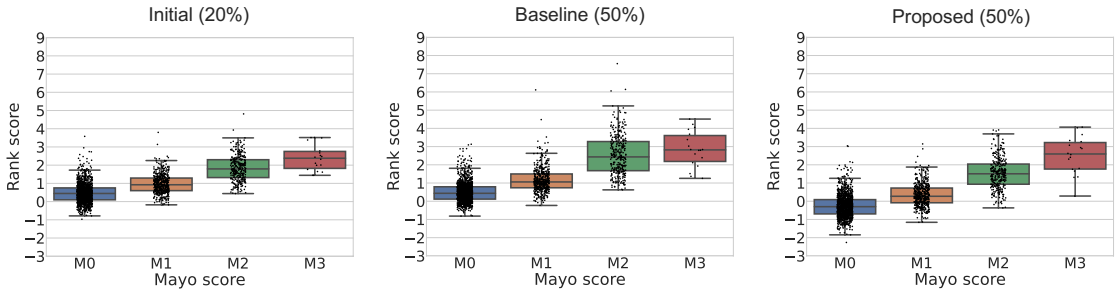


Figure 5-3: Box plots of estimated rank scores at each Mayo. Initial was measured under the initial condition at the labeling ratio of 20%. Performances of Baseline and Proposed were measured at the labeling ratio of 50%. If the distributions of each Mayo score have less overlap, the estimation can be considered reasonable.

Figure 5-3 shows box plots of the estimated rank scores of three methods at each Mayo score. Here, the vertical axis indicates the estimated rank score, and "Initial" indicates the method that trained the network using only the initial training data without iterations (20% of training data). In this plot, if the distributions of each Mayo score have less overlap, the estimation can be considered reasonable. In Initial and "Baseline," the score distributions in Mayo 2 and 3 significantly overlapped. In contrast, our method improved the overlap distributions. This indicates that the estimated rank scores were more correlated to the Mayo scores.

5.4.5 Relationship between Uncertainty and Class Imbalance

As described in Section 5.4.4, I considered that improvement by our method is because our uncertainty-based sampling mitigated the class imbalance. Therefore, I investigated the relationship between the uncertainty and the number of the Mayo labels of the sampled images from the training data during iterations.

Figure 5-4 shows the number of sampled images by uncertainty-based sampling at each class when the labeling ratio was 20% ($k = 0$), 35% ($k = 3$), and 50% ($k = K = 6$), where the vertical axis indicates the average numbers of five-fold cross-validation. The initial training data ($k = 0$) has a class imbalance by that in all the training data due to random sampling. In 35% and 50%, the sampled images

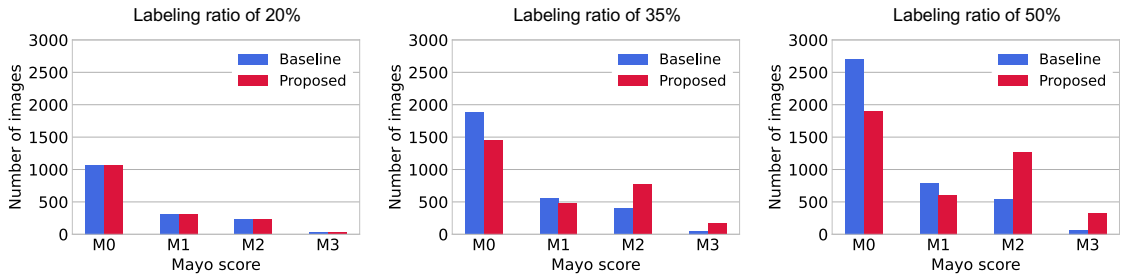


Figure 5-4: Class proportion of the accumulated sampled images at iteration $K = 0$ (labeling ratio is 20%), 3 (35%), and 6 (50%). When the labeling ratio was 20%, the class proportion was the same between Baseline and Proposed. The proposed method selected many samples of minority classes (Mayo 2 and 3) and mitigated the class imbalance problem.

by Baseline have a similar class imbalance to that in the initial training data. This class imbalance affected the performance improvements even though the number of training images increased, and thus the improvement was limited. In contrast, our uncertainty-based sampling selected many samples of the minority Mayo level; that is, the samples of Mayo 2 and 3 increased with iteration; therefore, the class imbalance was gradually mitigated with iteration. Consequently, our method improved relative label estimation accuracy despite the half size of training data compared to Baseline (all data).

Figure 5-5 shows the distributions of the model uncertainty of each class in training data. In "Baseline," the uncertainty of the minority classes (Mayo 2 and 3) was higher than the majority classes (Mayo 0 and 1). After six iterations of active learning (that is, the labeling ratio is 50%), the class imbalance was mitigated, as shown in Fig. 5-4, and the uncertainty of Mayo 2 and 3 was lower than those of "Baseline." This indicates that the uncertainty is correlated to the class imbalance; the fewer samples of a class are, the higher uncertainty is. Therefore, our uncertainty-based active learning, which selects the higher uncertainty samples as additional training data, can select many samples of the minority classes and consequently mitigate the class imbalance problem. This knowledge is useful for learning-to-rank tasks in medical image analysis since severe class imbalance problems often occur in medical images.

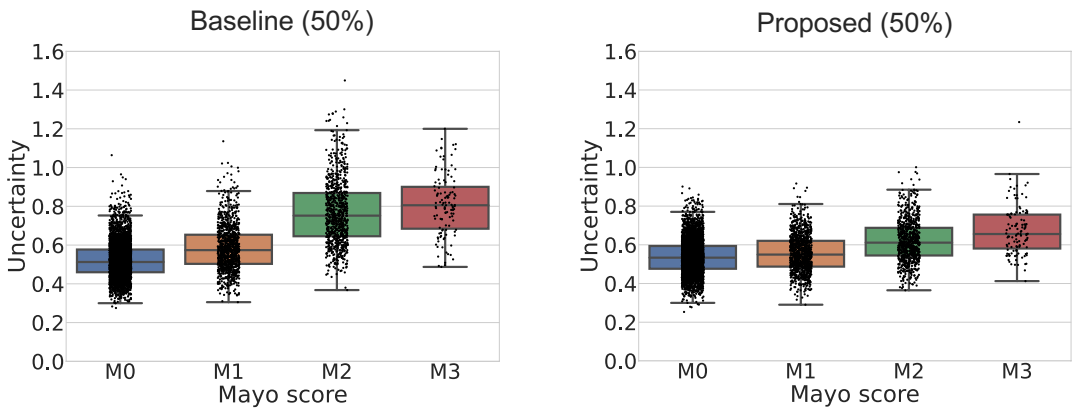


Figure 5-5: Box plots of model uncertainty in each class. Performance of Baseline and Proposed were measured at the labeling ratio of 50%. In Baseline, the uncertainty of minority classes (Mayo 2 and 3) was higher than that of majority classes. In Proposed, the uncertainty of Mayo 2 and 3 decreases since the class imbalance was mitigated.

5.4.6 Application by Estimated Rank Score

This section shows an example of the clinical applications of the proposed method. In UC severity diagnosis, endoscopic images are acquired in sequence while the endoscope is moved through the colon. In clinical, a doctor checks all the images to find the most severe images and diagnoses the severity of the UC. It is useful to show the estimated severity score along with the capturing order to facilitate this diagnosis process. Figure 5-6 shows three examples of the changes in the rank score (severity), where the horizontal axis indicates the capturing order in a sequence, and the vertical axis indicates the estimated rank score. These sequences were taken from different patients and were not used in the training data. In these results, the estimated rank scores (red) were similar to the Mayo scores (blue) that medical doctors annotated. Case 1 was a sequence of a patient with a mild disease level. The severity was low in the entire region. Cases 2 and 3 were sequences of patients with moderate and severe disease levels. In these examples, I can observe that the severe areas were biased in the order of the sequence; the severity was high at the beginning and the end of the sequence, but it was low in the middle of the sequence. Using this graph, medical

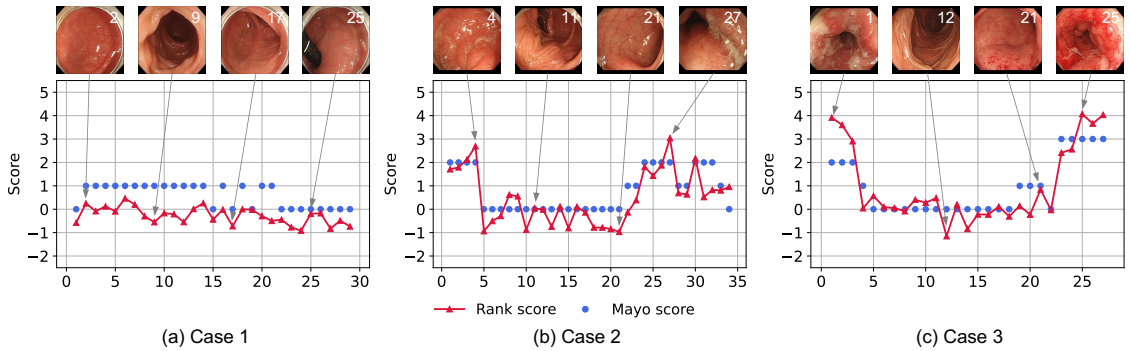


Figure 5-6: Examples of changes in the rank score along with the capturing order in a sequence. The horizontal axis indicates the capturing order, and the vertical axis indicates the estimated rank score (red) and the Mayo score (blue).

doctors can easily check if the disease is severe or not and find the severe areas and check them. In addition, it is also easy to show both cases before and after treatment and diagnose the recovery of the disease.

5.5 Conclusion

In this chapter, I proposed a deep Bayesian active-learning-to-rank for efficient relative annotation. The proposed method actively determines effective sample pairs for additional relative annotations by estimating the uncertainty using a Bayesian CNN. I first evaluated the accuracy and efficiency of the proposed method with an experiment about the correctness of the relative severity between a pair of images. The results indicate the usefulness of uncertainty-based active learning for selecting samples for better ranking. I also revealed that the proposed method selects minor but important samples and thus shows the robustness to class imbalance.

The limitation of the proposed method is that it provides rank scores instead of Mayo scores that medical experts are familiar with. If an application needs Mayo score-like rank scores, an additional calibration process is necessary. The proposed method is very general and applicable to other severity-level estimation tasks — this means more experiments on different image datasets are important for future tasks.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis proposed two learning-to-rank methods with relative annotation in medical image analysis. The first is a multi-task learning method that combines learning to rank and regression to calibrate rank scores to absolute labels. The second is active learning using a Bayesian CNN for learning to rank. These methods substantially reduced the annotation cost of labeling medical images that require to be performed by a medical expert. The performance of these methods was higher than that of conventional methods. These results indicated that learning to rank with relative annotation is more efficient than with absolute annotation in medical image analysis.

First, I used learn-to-rank with relative annotation for disease severity estimation to reduce annotation costs in medical image analysis. I proposed multi-task learning that combines learning to rank with relative annotation and regression with absolute annotation. The multi-task learning successfully calibrated the predictive value by learning to rank to the disease severity score. Experimental results show that the proposed method can achieve higher classification performance (accuracy and F1 score) with 1/10 of the annotation cost compared to conventional classification methods.

Second, I explored active learning using a Bayesian CNN to solve the problem of

increasing the number of pairs, which is one of the challenges of relative annotation. I proposed a deep Bayesian active-learning-to-rank for efficient relative annotation. The proposed method actively determines effective image pairs for additional relative annotation by estimating uncertainty using Bayesian CNN. The performance of the proposed method was evaluated through experiments on estimating the relative severity between image pairs. The results showed that uncertainty-based sampling could effectively select image pairs for relative severity estimation. I also showed that the proposed method is robust against class imbalance because it selects minor but important samples.

6.2 Future Work

- **Application of learning to rank and annotation for clinical practices:** Relative annotation, which is the annotation method used for learning to rank, is based on the intuitive judgment of the medical expert. Analysis of the annotations by this intuitive judgment will be necessary in the future. For example, it is necessary to examine the variability in annotation accuracy among multiple medical experts and to clarify the criteria for judgment. Furthermore, new medical image findings may be obtained from the criteria for judgment of relative annotation and total order.

- **Ranking tasks and domain shift in medical images:** The domain-shift issue is an important task in medical image analysis. For example, trained models on images taken with one device often fail to predict images taken with a different device. This issue means automatic prediction is impossible for patients examined at multiple hospitals. Therefore, exploring deep learning that can be used in multiple hospitals is important. On the other hand, domain-shift tasks related to learning to rank have yet to be sufficiently studied and should be explored in the future.

- **Noise problem of uncertainty-based sampling in medical images:** The uncertainty-based sampling selects samples with high learning effectiveness, but it also selects samples with noise. The reason is that the uncertainty is higher for samples with low similarity in the dataset. For endoscopic images, the images are selected with equipment in the image or at an angle that is not typical. Therefore, it is necessary to implement a mechanism to remove noise images in uncertainty-based sampling in the future.

Bibliography

[1] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *Proceedings of the Medical Imaging 2018: Image Processing*, vol. 10574, 2018, p. 105741M. [Online]. Available: https://doi.org/10.1117/12.2293971

[2] S. Xu, H. Wu, and R. Bie, "CXNet-m1: Anomaly Detection on Chest X-Rays With Image-Based Deep Learning," *IEEE Access*, vol. 7, pp. 4466–4477, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2885997

[3] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, and Y. Xia, "Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 879–890, 2021. [Online]. Available: https://doi.org/10.1109/TMI.2020.3040950

[4] D. Zhang, Y. Song, D. Liu, H. Jia, S. Liu, Y. Xia, H. Huang, and W. Cai, "Panoptic Segmentation with an End-to-End Cell R-CNN for Pathology Image Analysis," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018, pp. 237–244. [Online]. Available: https://doi.org/10.1007/978-3-030-00934-2_27

[5] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, "Adaptive Weighting Multi-Field-Of-View CNN for Semantic Segmentation in Pathology," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 597–12 606. [Online]. Available: https://doi.org/10.1109/CVPR.2019.01288

[6] R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017. [Online]. Available: https://doi.org/10.1016/j.ophtha.2017.02.008

[7] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative*

ophthalmology & visual science, vol. 57, pp. 5200–5206, 2016. [Online]. Available: https://doi.org/10.1167/iovs.16-19964

[8] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *European radiology*, vol. 31, no. 8, pp. 6096–6104, 2021. [Online]. Available: https://doi.org/10.1007/s00330-021-07715-1

[9] T.-Y. Liu, *Learning to Rank for Information Retrieval*. Springer, 2011.

[10] K. Crammer and Y. Singer, "Pranking with Ranking," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 14, 2001.

[11] F. C. Gey, "Inferring Probability of Relevance Using the Method of Logistic Regression," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 222–231. [Online]. Available: https://doi.org/10.1007/978-1-4471-2099-5_23

[12] R. Nallapati, "Discriminative Models for Information Retrieval," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 64–71. [Online]. Available: https://doi.org/10.1145/1008992.1009006

[13] R. Herbrich, T. Graepel, and K. Obermayer, "Large Margin Rank Boundaries for Ordinal Regression," in *Advances in Large Margin Classifiers*, 2000, pp. 115–132.

[14] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning (ICML)*, 2005, pp. 89–96. [Online]. Available: https://doi.org/10.1145/1102351.1102363

[15] L. Rigutini, T. Papini, M. Maggini, and F. Scarselli, "SortNet: learning to rank by a neural preference function," *IEEE Transactions on Neural Networks*, vol. 22, pp. 1368–1380, 2011. [Online]. Available: https://doi.org/10.1109/tnn.2011.2160875

[16] J. Xu and H. Li, "AdaRank: A Boosting Algorithm for Information Retrieval," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 391–398. [Online]. Available: https://doi.org/10.1145/1277741.1277809

[17] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to Rank: From Pairwise Approach to Listwise Approach," in *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007, pp. 129–136. [Online]. Available: https://doi.org/10.1145/1273496.1273513

[18] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise Approach to Learning to Rank: Theory and Algorithm," in *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008, pp. 1192–1199. [Online]. Available: https://doi.org/10.1145/1390156.1390306

[19] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal Regression Methods: Survey and Experimental Study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016. [Online]. Available: https://doi.org/10.1109/TKDE.2015.2457911

[20] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2708503

[21] X. Jiang, L. Shen, L. Yu, M. Jiang, and G. Feng, "No-reference screen content image quality assessment based on multi-region features," *Neurocomputing*, vol. 386, pp. 30–41, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.12.027

[22] J. Yan, S. Lin, and S. B. Kang, "A learning-to-rank approach for image color enhancement," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2987–2994. [Online]. Available: https://doi.org/10.1109/CVPR.2014.382

[23] N. Ma, A. Volkov, A. Livshits, P. Pietrusinski, H. Hu, and M. Bolin, "An Universal Image Attractiveness Ranking Framework," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 657–665. [Online]. Available: https://doi.org/10.1109/WACV.2019.00075

[24] Y. Murata and Y. Dobashi, "Automatic image enhancement taking into account user preference," in *Proceedings of the 2019 International Conference on Cyberworlds (CW)*, 2019, pp. 374–377. [Online]. Available: https://doi.org/10.1109/CW.2019.00070

[25] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to Rank for Blind Image Quality Assessment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2275–2290, 2015. [Online]. Available: https://doi.org/10.1109/TNNLS.2014.2377181

[26] Y. Niu, D. Huang, Y. Shi, and X. Ke, "Siamese-Network-Based Learning to Rank for No-Reference 2D and 3D Image Quality Assessment," *IEEE Access*, vol. 7, pp. 101 583–101 595, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2930707

[27] W. Huang, K. L. Chan, H. Li, J. H. Lim, J. Liu, and T. Y. Wong, "A computer assisted method for nuclear cataract grading from slit-lamp images using ranking," *IEEE Transactions on Medical Imaging*, vol. 30, no. 1, pp. 94–107, 2011. [Online]. Available: https://doi.org/10.1109/TMI.2010.2062197

[28] F. Pedregosa, E. Cauvet, G. Varoquaux, C. Pallier, B. Thirion, and A. Gramfort, "Learning to Rank from Medical Imaging Data," in *Proceedings of the Machine Learning in Medical Imaging (MLMI)*, 2012, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-642-35428-1_29

[29] B. Peng, X. Yao, S. L. Risacher, A. J. Saykin, L. Shen, and X. Ning, "Prioritization of Cognitive Assessments in Alzheimer's Disease via Learning to Rank using Brain Morphometric Data," in *Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, pp. 1–4. [Online]. Available: https://doi.org/10.1109/BHI.2019.8834618

[30] J. Lyu, S. H. Ling, S. Banerjee, J. J. Zheng, K.-L. Lai, D. Yang, Y.-P. Zheng, and S. Su, "3D Ultrasound Spine Image Selection Using Convolution Learning-to-Rank Algorithm," in *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 4799–4802. [Online]. Available: https://doi.org/10.1109/EMBC.2019.8857182

[31] G. Adîr, V. Adîr, and N. E. Pascu, "Logo Design and the Corporate Identity," *Procedia - Social and Behavioral Sciences*, vol. 51, pp. 650–654, 2012. [Online]. Available: https://doi.org/10.1016/j.sbspro.2012.08.218

[32] W. Zhao, "A brief analysis on the new trend of logo design in the digital information era," in *Proceedings of the 2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME)*, 2017. [Online]. Available: https://doi.org/10.2991/essaeme-17.2017.356

[33] A. Sundar and T. J. Noseworthy, "Place the Logo High or Low? Using Conceptual Metaphors of Power in Packaging Design," *Journal of Marketing*, vol. 78, no. 5, pp. 138–151, 2014. [Online]. Available: https://doi.org/10.1509/jm.13.0253

[34] J. Luffarelli, A. Stamatogiannakis, and H. Yang, "The Visual Asymmetry Effect: An Interplay of Logo Design and Brand Personality on Brand Equity," *Journal of marketing research*, vol. 56, no. 1, pp. 89–103, 2019. [Online]. Available: https://doi.org/10.1177/0022243718820548

[35] J. Luffarelli, M. Mukesh, and A. Mahmood, "Let the Logo Do the Talking: The Influence of Logo Descriptiveness on Brand Equity," *Journal*

of Marketing Research, vol. 56, no. 5, pp. 862–878, 2019. [Online]. Available: https://doi.org/10.1177/0022243719845000

[36] J. Neumann, H. Samet, and A. Soffer, "Integration of local and global shape analysis for logo classification," *Pattern Recognition Letters*, vol. 23, no. 12, pp. 1449–1457, 2002. [Online]. Available: https://doi.org/10.1016/S0167-8655(02)00105-8

[37] H. Su, S. Gong, and X. Zhu, "WebLogo-2M: Scalable Logo Detection by Deep Learning from the Web," in *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 270–279. [Online]. Available: https://doi.org/10.1109/ICCVW.2017.41

[38] A. Sage, R. Timofte, E. Agustsson, and L. V. Gool, "Logo Synthesis and Manipulation with Clustered Generative Adversarial Networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5879–5888. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00616

[39] J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, H. Wang, and S. Jiang, "Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Apr. 2020, pp. 6194–6201. [Online]. Available: https://doi.org/10.1609/aaai.v34i04.6085

[40] T. Karamatsu, D. Suehiro, and S. Uchida, "Logo Design Analysis by Ranking," in *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1482–1487. [Online]. Available: https://doi.org/10.1109/ICDAR.2019.00238

[41] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A Deep Look into neural ranking models for information retrieval," *Information Processing & Management*, vol. 57, no. 6, p. 102067, 2020. [Online]. Available: https://doi.org/10.1016/j.ipm.2019.102067

[42] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose, "Spatio-Temporal Analysis of Topic Popularity in Twitter," *arXiv preprint arXiv:1111.2904*, 2011. [Online]. Available: https://doi.org/10.48550/arXiv.1111.2904

[43] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an Influencer: Quantifying Influence on Twitter," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 2011, pp. 65–74. [Online]. Available: https://doi.org/10.1145/1935826.1935845

[44] K. Lerman, R. Ghosh, and T. Surachawala, "Social Contagion: An Empirical Study of Information Spread on Digg and Twitter Follower Graphs," *arXiv preprint arXiv:1202.3162*, 2012. [Online]. Available: https://doi.org/10.48550/arXiv.1202.3162

[45] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao, "Follow the green: Growth and dynamics in twitter follower markets," in *Proceedings of the 2013 Conference on Internet Measurement Conference*, 2013, pp. 163–176. [Online]. Available: https://doi.org/10.1145/2504730.2504731

[46] K. W. Schroeder, W. J. Tremaine, and D. M. Ilstrup, "Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis," *The New England Journal of Medicine*, vol. 317, no. 26, pp. 1625–1629, 1987. [Online]. Available: https://doi.org/10.1056/nejm198712243172603

[47] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using Ranking-CNN for Age Estimation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 742–751. [Online]. Available: https://doi.org/10.1109/CVPR.2017.86

[48] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A Ranking Approach for Human Ages Estimation Based on Face Images," in *Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3396–3399. [Online]. Available: https://doi.org/10.1109/ICPR.2010.829

[49] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4920–4928. [Online]. Available: https://doi.org/10.1109/CVPR.2016.532

[50] B. Liu, Y. Zhang, M. Chu, X. Bai, and F. Zhou, "Bone age assessment based on rank-monotonicity enhanced ranking CNN," *IEEE Access*, vol. 7, pp. 120 976–120 983, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2937341

[51] X. Liu, Y. Zou, Y. Song, C. Yang, J. You, and B. V. K. V. Kumar, "Ordinal Regression with Neuron Stick-Breaking for Medical Diagnosis," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2019, pp. 335–344. [Online]. Available: https://doi.org/10.1007/978-3-030-11024-6_23

[52] T. J. Jun, Y. Eom, D. Kim, C. Kim, J.-H. Park, H. M. Nguyen, Y.-H. Kim, and D. Kim, "TRk-CNN: Transferable Ranking-CNN for image classification of glaucoma, glaucoma suspect, and normal eyes," *Expert Systems with Applications*, vol. 182, p. 115211, 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.115211

[53] B. Wu, X. Sun, L. Hu, and Y. Wang, "Learning With Unsure Data for Medical Image Diagnosis," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10 589–10 598. [Online]. Available: https://doi.org/10.1109/ICCV.2019.01069

[54] L. Li and H.-t. Lin, "Ordinal Regression by Extended Binary Classification," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 19, 2006.

[55] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of the 2011 International Conference on Computer Vision (ICCV)*, 2011, pp. 503–510. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126281

[56] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim, "Deep Relative Attributes," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1832–1842, 2016. [Online]. Available: https://doi.org/10.1109/TMM.2016.2582379

[57] F. Xiao and Y. J. Lee, "Discovering the spatial extent of relative attributes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015, pp. 1458–1466.

[58] K. K. Singh and Y. J. Lee, "End-to-End Localization and Ranking for Relative Attributes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 753–769. [Online]. Available: https://doi.org/10.1007/978-3-319-46466-4_45

[59] J. Kalpathy-Cramer, J. P. Campbell, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J. D. Reynolds, K. Hutcheson, M. J. Shapiro, M. X. Repka *et al.*, "Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis," *Ophthalmology*, vol. 123, no. 11, pp. 2345–2351, 2016. [Online]. Available: https://doi.org/10.1016/j.ophtha.2016.07.020

[60] G. Saibro, M. Diana, B. Sauer, J. Marescaux, A. Hostettler, and T. Collins, "Automatic Detection of Steatosis in Ultrasound Images with Comparative Visual Labeling," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022, p. 408–418. [Online]. Available: https://doi.org/10.1007/978-3-031-16437-8_39

[61] F. Hirai and T. Matsui, "A critical review of endoscopic indices in ulcerative colitis: inter-observer variation of the endoscopic index," *Clinical Journal of Gastroenterology*, vol. 1, no. 2, pp. 40–45, 2008. [Online]. Available: https://doi.org/10.1007/s12328-008-0018-z

[62] J. H. Lee, Y. J. Kim, Y. W. Kim, S. Park, Y. i. Choi, Y. J. Kim, D. K. Park, K. G. Kim, and J. W. Chung, "Spotting malignancies from gastric endoscopic images using deep learning," *Surgical Endoscopy*, vol. 33, no. 11, pp. 3790–3797, 2019. [Online]. Available: https://doi.org/10.1007/s00464-019-06677-2

[63] Y. Zhu, Q. C. Wang, M. D. Xu, Z. Zhang, J. Cheng, Y. S. Zhong, Y. Q. Zhang, W. F. Chen, L. Q. Yao, P. H. Zhou, and Q. L. Li, "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy," *Gastrointestinal Endoscopy*, vol. 89, no. 4, pp. 806–815, 2019. [Online]. Available: https://doi.org/10.1016/j.gie.2018.11.011

[64] B.-J. Cho, C. S. Bang, S. W. Park, Y. J. Yang, S. I. Seo, H. Lim, W. G. Shin, J. T. Hong, Y. T. Yoo, S. H. Hong *et al.*, "Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network," *Endoscopy*, vol. 51, no. 12, pp. 1121–1129, 2019. [Online]. Available: https://doi.org/10.1055/a-0981-6133

[65] X. Zhang, W. Hu, F. Chen, J. Liu, Y. Yang, L. Wang, H. Duan, and J. Si, "Gastric precancerous diseases classification using CNN with a concise model," *PLoS ONE*, vol. 12, no. 9, pp. 1–10, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0185508

[66] X. Liu, C. Wang, J. Bai, and G. Liao, "Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images," *Neurocomputing*, vol. 392, pp. 253–267, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.10.100

[67] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, Y. Takemura, K. Onji, R. Miyaki, and S. Tanaka, "Computer-aided colorectal tumor classification in NBI endoscopy: Using local features," *Medical Image Analysis*, vol. 17, no. 1, pp. 78–100, 2013. [Online]. Available: https://doi.org/10.1016/j.media.2012.08.003

[68] K. Choi, S. J. Choi, and E. S. Kim, "Computer-Aided Diagonosis for Colorectal Cancer using Deep Learning with Visual Explanations," in *2020 42nd annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 1156–1159. [Online]. Available: https://doi.org/10.1109/EMBC44109.2020.9176653

[69] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic Polyp Classification with Convolutional Neural Networks," in *Proceedings of the 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, 2016, pp. 253–258. [Online]. Available: https://doi.org/10.1109/CBMS.2016.39

[70] R. Zhang, Y. Zheng, C. C. Poon, D. Shen, and J. Y. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern recognition*, vol. 83, pp. 209–219, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2018.05.026

[71] M. Liu, J. Jiang, and Z. Wang, "Colonic Polyp Detection in Endoscopic Videos With Single Shot Detection Based Deep Convolutional Neural Network," *IEEE Access*, vol. 7, pp. 75 058–75 066, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2921027

[72] Y. Sakai, S. Takemoto, K. Hori, M. Nishimura, H. Ikematsu, T. Yano, and H. Yokota, "Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network," in *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 4138–4141. [Online]. Available: https://doi.org/10.1109/EMBC.2018.8513274

[73] L. Wu, W. Zhou, X. Wan, J. Zhang, L. Shen, S. Hu, Q. Ding, G. Mu, A. Yin, X. Huang *et al.*, "A deep neural network improves endoscopic detection of early gastric cancer without blind spots," *Endoscopy*, vol. 51, no. 06, pp. 522–531, 2019. [Online]. Available: https://doi.org/10.1055/a-0855-3532

[74] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily, "Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images," *Sensors*, vol. 19, no. 6, 2019. [Online]. Available: https://doi.org/10.3390/s19061265

[75] T. Aoki, A. Yamada, K. Aoyama, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, S. Tanaka, K. Koike, and T. Tada, "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 357–363, 2019. [Online]. Available: https://doi.org/10.1016/j.gie.2018.10.027

[76] M. K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, and K. Mori, "Automatic detection of informative frames from wireless capsule endoscopy images," *Medical Image Analysis*, vol. 14, no. 3, pp. 449–470, 2010. [Online]. Available: https://doi.org/10.1016/j.media.2009.12.001

[77] A. Dahal, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen, "Detection of ulcerative colitis severity in colonoscopy video frames," in *Proceedings of the 2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/CBMI.2015.7153617

[78] A. Alammari, A. R. Islam, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Classification of Ulcerative Colitis Severity in Colonoscopy Videos Using CNN," in *Proceedings of the 9th International Conference on Information Management and Engineering (ICIME)*, 2017, p. 139–144. [Online]. Available: https://doi.org/10.1145/3149572.3149613

[79] R. W. Stidham, W. Liu, S. Bishu, M. D. Rice, P. D. Higgins, J. Zhu, B. K. Nallamothu, and A. K. Waljee, "Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis," *JAMA network open*, vol. 2, no. 5, p. e193963, 2019. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.3963

[80] H. Yao, K. Najarian, J. Gryak, S. Bishu, M. D. Rice, A. K. Waljee, H. J. Wilkins, and R. W. Stidham, "Fully automated endoscopic disease activity assessment in ulcerative colitis," *Gastrointestinal Endoscopy*, vol. 93, no. 3, pp. 728–736.e1, 2021. [Online]. Available: https://doi.org/10.1016/j.gie.2020.08.011

[81] K. Gottlieb, J. Requa, W. Karnes, R. Chandra Gudivada, J. Shen, E. Rael, V. Arora, T. Dao, A. Ninh, and J. McGill, "Central reading of ulcerative colitis clinical trial videos using neural networks," *Gastroenterology*, vol. 160, no. 3, pp. 710–719.e2, 2021. [Online]. Available: https://doi.org/10.1053/j.gastro.2020.10.024

[82] Y. Fan, R. Mu, H. Xu, C. Xie, Y. Zhang, L. Liu, L. Wang, H. Shi, Y. Hu, J. Ren, J. Qin, L. Wang, and S. Cai, "Novel deep learning–based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis," *Gastrointestinal Endoscopy*, 2022. [Online]. Available: https://doi.org/10.1016/j.gie.2022.08.015

[83] K. Takenaka, K. Ohtsuka, T. Fujii, M. Negi, K. Suzuki, H. Shimizu, S. Oshima, S. Akiyama, M. Motobayashi, M. Nagahori, E. Saito, K. Matsuoka, and M. Watanabe, "Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis," *Gastroenterology*, vol. 158, no. 8, pp. 2150–2157, 2020. [Online]. Available: https://doi.org/10.1053/j.gastro.2020.02.012

[84] X. Luo, J. Zhang, Z. Li, and R. Yang, "Diagnosis of ulcerative colitis from endoscopic images based on deep learning," *Biomedical Signal Processing and Control*, vol. 73, p. 103443, 2022. [Online]. Available: https://doi.org/10.1016/j.bspc.2021.103443

[85] M. Turan and F. Durmus, "UC-NfNet: Deep learning-enabled assessment of ulcerative colitis from colonoscopy images," *Medical Image Analysis*, vol. 82, p. 102587, 2022. [Online]. Available: https://doi.org/10.1016/j.media.2022.102587

[86] E. Schwab, G. O. Cula, K. Standish, S. S. F. Yip, A. Stojmirovic, L. Ghanem, and C. Chehoud, "Automatic estimation of ulcerative colitis severity from endoscopy videos using ordinal multi-instance learning," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–9, 2021. [Online]. Available: https://doi.org/10.1080/21681163.2021.1997644

[87] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. [Online]. Available: https://doi.org/10.1109/CVPR.2017.243

[88] Y. Yuan, W. Qin, B. Ibragimov, B. Han, and L. Xing, "RIIS-DenseNet: Rotation-Invariant and Image Similarity Constrained Densely Connected Convolutional Network for Polyp Detection," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018, pp. 620–628. [Online]. Available: https://doi.org/10.1007/978-3-030-00934-2_69

[89] R. Bise, K. Abe, H. Hayashi, K. Tanaka, and S. Uchida, "Efficient Soft-Constrained Clustering for Group-Based Labeling," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019, pp. 421–430. [Online]. Available: https://doi.org/10.1007/978-3-030-32254-0_47

[90] D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 7, 1994.

[91] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 1183–1192.

[92] O. Sener and S. Savarese, "Active Learning for Convolutional Neural Networks: A Core-Set Approach," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[93] H. Zheng, L. Yang, J. Chen, J. Han, Y. Zhang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "Biomedical image segmentation via representative annotation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5901–5908. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33015901

[94] X. Wu, C. Chen, M. Zhong, and J. Wang, "HAL: Hybrid active learning for efficient labeling in medical domain," *Neurocomputing*, vol. 456, pp. 563–572, 2021. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.10.115

[95] Q. Jin, M. Yuan, Q. Qiao, and Z. Song, "One-shot active learning for image segmentation via contrastive learning and diversity-based sampling," *Knowledge-Based Systems*, vol. 241, p. 108278, 2022. [Online]. Available: https://doi.org/10.1016/j.knosys.2022.108278

[96] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-Effective Active Learning for Deep Image Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2017. [Online]. Available: https://doi.org/10.1109/TCSVT.2016.2589879

[97] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017, pp. 399–407. [Online]. Available: https://doi.org/10.1007/978-3-319-66179-7_46

[98] S. Wen, T. Kurc, L. Hou, J. Saltz, R. Gupta, R. Batiste, T. Zhao, V. Nguyen, D. Samaras, and W. Zhu, "Comparison of Different Classifiers with Active Learning to Support Quality Control in Nucleus Segmentation in Pathology Images," *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, vol. 2017, pp. 227–236, 2018.

[99] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation," *Medical Image Analysis*, vol. 59, p. 101557, 2020. [Online]. Available: https://doi.org/10.1016/j.media.2019.101557

[100] W. Li, J. Li, Z. Wang, J. Polson, A. E. Sisk, D. P. Sajed, W. Speier, and C. W. Arnold, "PathAL: An Active Learning Framework for Histopathology Image Analysis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1176–1187, 2022. [Online]. Available: https://doi.org/10.1109/TMI.2021.3135002

[101] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 113–127, 2015. [Online]. Available: https://doi.org/10.1007/s11263-014-0781-x

[102] V. Nath, D. Yang, B. A. Landman, D. Xu, and H. R. Roth, "Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2534–2547, 2021. [Online]. Available: https://doi.org/10.1109/TMI.2020.3048055

[103] K. Margatina, G. Vernikos, L. Barrault, and N. Aletras, "Active Learning by Acquiring Contrastive Examples," in *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, nov 2021, pp. 650–663. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.51

[104] Q. Wang, W. Wu, Y. Qi, and Y. Zhao, "Deep Bayesian Active Learning for Learning to Rank: A Case Study in Answer Selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5251–5262, 2022. [Online]. Available: https://doi.org/10.1109/TKDE.2021.3056894

[105] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, vol. 48, 2016, pp. 1050–1059. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[106] T. Kadota, K. Abe, R. Bise, T. Kawamura, N. Sakiyama, K. Tanaka, and S. Uchida, "Automatic Estimation of Ulcerative Colitis Severity by Learning to Rank With Calibration," *IEEE Access*, vol. 10, pp. 25 688–25 695, 2022. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3155769

[107] Q. Xu, Z. Yang, Z. Chen, Y. Jiang, X. Cao, Y. Yao, and Q. Huang, "Deep Partial Rank Aggregation for Personalized Attributes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, May 2021, pp. 678–688. [Online]. Available: https://doi.org/10.1609/aaai.v35i1.16148

[108] Y. You, C. Lu, W. Wang, and C.-K. Tang, "Relative CNN-RNN: Learning Relative Atmospheric Visibility From Images," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 45–55, 2019. [Online]. Available: https://doi.org/10.1109/TIP.2018.2857219