

Top-Rank Learning with Neural Networks

鄭, 顔

<https://hdl.handle.net/2324/4784643>

出版情報 : Kyushu University, 2021, 博士 (学術), 課程博士
バージョン :
権利関係 :



Top-Rank Learning with Neural Networks

Yan Zheng

Department of Advanced Information Technology

Doctor of Philosophy

KYUSHU UNIVERSITY

February 2022

Top-Rank Learning with Neural Networks

Yan Zheng

Department of Advanced Information Technology

February 2022

Doctor of Philosophy

Abstract

Top-rank learning aims to learn a real-valued ranking function to maximize the number of *absolute top* samples. These samples are the most reliable positive samples that are ranked higher than the top-ranked negative sample. Therefore, top-rank learning is beneficial for real applications that require reliable decisions. In addition, it also inherits the merits of bipartite ranking that not only provides the appropriate ranking for samples but also theoretically guarantees its robustness to data imbalance problems. In this thesis, we first propose to apply a bipartite ranking method (RankSVM) for solving the data imbalance problem in the signature verification task as preliminary research of top-rank learning. The experimental results demonstrate the effectiveness and superiority of RankSVM on various signature verification datasets, and they also present the effect of feature space for RankSVM statistically. Then, we improve top-rank learning and apply them to real applications. Specifically, we try to remove the following two limitations of conventional top-rank learning: (1) Most tasks have been formulated as linear or kernel-based problems and are thus limited in coping with complicated tasks; (2) It is sensitive to outliers of the negative samples. To tackle the first limitation, we propose a novel deep learning model—Top-rank Convolutional Neural Networks (TopRank CNN) that combines top-rank learning with representation learning. Since the original objective function of top-rank learning suffers from overfitting, we employ the p -norm relaxation of the initial loss function in the proposed method. We prove the usefulness of TopRank CNN experimentally with medical diagnosis tasks that require reliable decisions and robustness to the unbalanced condition. For the second limitation, we propose a robust top-rank learning algorithm with an unsupervised outlier estimation technique called the local outlier factor (LOF). Introduction of LOF can help weaken the effect of the negative outliers and thus increase the stability of the learned ranking function. Moreover, we combine robust top-rank learning with representation learning by a deep neural network (DNN). Experiments on artificial datasets and a medical image dataset demonstrate the robustness of the proposed method to outliers.

Thesis Supervisor: Seiichi Uchida

Title: Professor

Acknowledgments

I am very cherish this opportunity to express my gratitude to all those who helped me during my Ph.D. It was the most precious time in my life because of everyone and the things I met in the past three and half years. My gains are not only on the improvement of researches and significantly impacting my future life.

First of all, I would like to thank my supervisor, Prof. Seiichi Uchida, for his continuous support, careful guidance and help during my Ph.D life. I still remember the first time I met Prof. Uchida; it was also the first time I participated in the "interview" in English, his warm care and kindly guidance made me not nervous at all. Prof. Uchida is also my senior, friend, and colleague. Most time, he is like a father to me. When I encounter difficulties or make mistakes, Prof. Uchida always encourages me and gives me constructive comments; When I make progress, he always gives me a big thumbs up and feels happy for me. Prof. Uchida also cares about the emotions of the students. Prof. Uchida is always busy, but he still carefully prepared a meeting and slide to comfort me and discuss the issue of my graduation postponement due to the delay in publishing the Journal paper. I am very lucky that I have such a responsible, excellent, and nice supervisor. His wide knowledge, strong research enthusiasm, and hard-working attitude have inspired me during my life and study and will have a profound effect on my academic career forever. Without his support, guidance, patience, and encouragement, I could not have imagined how I could finish my Ph.D. I would like to thank my distinguished supervisor, Prof. Seiichi Uchida, again for his contribution during my Ph.D. life. I will never forget the story Prof. Uchida told me about how to follow "知彼知己，百战不殆" in researches and conform it for my future careers.

Next, I would like to thank Assistant Prof. Daiki Suehiro and my predecessor, Associate Prof. Yuchen Zheng (Now, he is in Shihezi University in China) for their kindly care during these three and half years. Prof. Suehiro gives me so much guidance for my researches. His well-knit theoretical knowledge and rigorous attitude

to research inspired me a lot for my future researches. Yuchen is my predecessor from my master, who is also my good friend and colleague. He gave me many useful advice and guidance both in my research and daily life. When I encounter some difficulties, he always comforts me when I am feeling upset. Both Prof. Suehiro and Yuchen are excellent colleagues and partners during my research life. They inspired me a lot and always helps me revise my paper both in technique and language. I wasn't able to successfully complete my Ph.D. program without your help. Thank you for your teaching and help.

Then, I would like to thank Prof. Wataru Ohyama (Now, he is in Intelligent Media Processing Laboratory, Saitama Institute of Technology) and Assistant Prof. Brian Kenji Iwana, who are my excellent colleagues during my Ph.D. researches. They help me a lot about my researches. Prof. Iwana also helps me schedule my Ph.D. life and reminds me of what should I do at the right time and places.

Also, I would like to thank my laboratory, Human Interface Laboratory at Kyushu University. As a member of our laboratory, I feel so lucky and happy and cherish the time spent with each lab member. I would like to thank Associate Prof. Ryoma Bise, Assistant Prof. Hideaki Hayashi. They care about me a lot and give many valuable comments on my research. I would like to thank Secretary Mrs. Mika Maeda and Maki Fukutomi, and all Doctor, Master, Bachelor students in our lab. With all your help, I could intently focus on my research. I appreciate all the people who help me in our lab. I love our lab (home) forever.

Then, I would like to thank my family (my parents and my little sister Zi Zheng), who always accompanied me and stood by me. During these three and half years, I can not take care of my parents because I need to complete my Ph.D. in a foreign country. However, they never complain about that and always encourage me. I am very grateful to my sister to take care of my parents and handle the chores at home. Even though she is younger than me, she is always optimistic and comforts me when I feel upset. Really thanks to my family. I also want to thank my friends,

especially the Chinese friends in Fukuoka, such as my junior of our lab, Xiaotong Ji, My neighborhoods Kang Zhang, Xiao Ye, Wenbo Li, and My Japanese class partner Lin Hao, my previous roommates Xueru Zhang and Wanju Wei. They give me so much support and help in my Ph.D. life. Thanks again for your kindly cares.

Finally, I would like to thank Kyushu University and CSC (China Scholarship Council) for sponsoring my Ph.D. with a fully-funded scholarship. Since 2020, we have been affected by COVID-19. I believe that in the near future, the virus will be defeated, and the healthy world will come back with the efforts of all the people. Wish the spring arrives, all the good things will arrive in time.

Contents

1	Introduction	21
1.1	Background	21
1.2	Motivation	23
1.2.1	Usefulness of Bipartite Ranking	23
1.2.2	Properties of Top-Rank Learning	23
1.2.3	Limitations in the Conventional Top-Rank Learning	25
1.3	Thesis Objectives and Solutions	26
1.3.1	A Preliminary Research of Top-Rank Learning	26
1.3.2	Proposal of TopRank CNN	27
1.3.3	Proposal of Robust Top-Rank Learning	28
1.4	Novelty and Contributions	29
1.5	Thesis Structure	30
2	A General Description of Learning to Rank	33
2.1	Learning to Rank	33
2.2	RankSVM	34
2.3	Top-Rank Learning	36
3	RankSVM for Offline Signature Verification	39
3.1	Purpose	39
3.2	Related Work	40

3.2.1	Deep Learning methods for Signature Verification	40
3.2.2	Imbalanced Problem in Signature Verification	41
3.3	Utilizing the RankSVM for Offline Signature Verification	42
3.3.1	Implementation of RankSVM for Skilled Forgery Attacks . . .	42
3.3.2	Implementation of RankSVM for Signature Retrieval	43
3.4	Experiments for Skilled Forgery Attacks	43
3.4.1	Evaluation Metrics	44
3.4.2	Experimental Result with Bitmap	45
3.4.3	Experimental Results with Deep Feature	46
3.5	Experiments in Signature Retrieval	47
3.5.1	Quantitative Evaluation and Qualitative Evaluation	47
3.6	Summary	48
4	TopRank CNN for Medical Image Analysis	51
4.1	Purpose	51
4.2	Related Work	52
4.2.1	Learning to Rank with Deep Neural Networks	52
4.2.2	Top-Rank Learning with Deep Neural Networks	53
4.3	Top-Rank Convolutional Neural Networks	54
4.3.1	p-norm Relaxation	54
4.3.2	TopRank CNN	55
4.4	Experiment on an Unbalanced CIFAR-10 Dataset	56
4.4.1	Dataset	56
4.4.2	Comparative Methods	57
4.4.3	TopRank CNN Architecture and Hyper-parameters	58
4.4.4	Results on the Unbalanced CIFAR-10 Dataset	59
4.5	Experiment on a Medical Image Diagnosis Task	61
4.5.1	Dataset	62
4.5.2	Comparative Methods	63

<i>CONTENTS</i>	11
4.5.3 Results on the Medical Image Diagnosis Task	64
4.5.4 Results on Extremely Unbalanced Data	65
4.6 Summary	68
5 Robust Top-rank learning for Medical Image Analysis	71
5.1 Purpose	71
5.2 Related Work about Outlier Estimation	72
5.3 Robust Top-Rank Deep Neural Network	73
5.3.1 Top-Rank Learning Robust to Outlier	73
5.3.2 Combination with Representation Learning	74
5.4 Experiment on Artificial Datasets	75
5.5 Experiment on Medical Image Diagnosis	78
5.5.1 Dataset	78
5.5.2 Model Setup	79
5.5.3 Comparative Methods	79
5.5.4 Quantitative Evaluation by Pos@Top	81
5.5.5 Quantitative Evaluation of Ranking stability	81
5.6 Summary	82
6 Conclusion and Future Work	85
6.1 Conclusion	85
6.2 Future Work	87

List of Figures

1-1	Overview of (a) bipartite ranking, (b) top-rank learning, and (c) TopRank CNN (the proposed method). The green dots denote the positive samples and the orange dots denote the negative samples.	22
1-2	Relationship between the rankings and their ROC.	24
1-3	(a) Top-rank learning, and (b) its weakness to the negative outlier in the training set. In (b), the negative outlier becomes the top-ranked negative and, consequently, Pos@Top becomes much less than (a). Moreover, the ranking function becomes very different from (a) just by the single outlier. (c) The proposed method can weaken the effect of the outlier and get a more suitable ranking function like (a). Note that this figure does not focus on representation learning (i.e., feature extraction), although the proposed method does it along with top-rank learning in an end-to-end manner.	25

2-1	Overview of RankSVM. (a) Sample distributions of the positive class (genuine) and the negative class (forgery). The orange dots denote genuinely and the blue dots denote the forgery. (b) Result of standard SVM for classification. In the worst case, a forgery sample becomes the sample with the largest likelihood as the genuine, although it can minimize the classification error. (c) Result of RankSVM, which can give the best pattern order, or equivalently, maximize the AUC. (d) If positive and negative classes are linearly separable, standard SVM and RankSVM give the same result.	35
3-1	Signature examples from GPDS dataset.	44
3-2	(a) The DET curve of RankSVM and the standard SVM with bitmap on GPDS-150 dataset. (b) The DET curve of RankSVM and the standard SVM with deep features on GPDS-960 dataset.	45
3-3	The ROC curve of RankSVM and standard SVM on GPDS-150, GPDS-300, GPDS-600, GPDS-1000 datasets.	48
3-4	Top-5 signatures for RankSVM and Standard SVM on GPDS-150 dataset. Each two rows denote one user's ranked signatures; the top and bottom rows show the top-5 results of RankSVM and the standard SVM, respectively. Images in the black box and the red box denote the correct genuine signatures and the wrong signatures, respectively.	49
4-1	The architecture of TopRank CNN.	55
4-2	Examples of positive and negative classes from the unbalanced CIFAR-10 dataset.	57

4-3	t-SNE visualizations of the training, validation, and test sets of the unbalanced CIFAR-10 dataset. For each method, the upper row shows positive (green and red; red denotes <i>the absolute top</i>) and negative (orange) sample distribution and the lower shows their ranking scores (darker color is higher). For the classification methods, such as CNN_{soft} , their soft-max values are used as the ranking scores. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)	61
4-4	Examples from the DDSM Mammography dataset. The upper row shows the masses or calcification images (positive) and the bottom row shows the normal images (negative).	62
4-5	(a) ROC of TopRank CNN (ours) and comparative methods on the test set of the DDSM mammography dataset. (b) The leftmost vertical part of (a).	64
4-6	Top-10 ranked samples in the test set by the TopRank CNN (ours). All of these samples are abnormal (positive) images according to their ground-truth.	65
4-7	t-SNE visualization of the training, validation, and test sets of the DDSM Mammography dataset. For each method, the upper row shows positive (green and red; red denotes <i>the absolute top</i>) and negative (orange) sample distribution and the lower shows their ranking scores (darker color is higher).(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)	66
4-8	Pos@Top of TopRank CNN (ours) with different p on the validation set of the DDSM Mammography dataset.	67

4-9	(a) ROC of TopRank CNN (ours) and comparative methods on the test set when the extremely unbalanced DDSM mammography data is used for training. (b) The leftmost part of (a).	67
4-10	(a) The training loss curve of the proposed TopRank CNN with $p=8$ when the extremely unbalanced DDSM Mammography training samples (whose the ratio of positive and negative samples is 1 : 97) are used. (b) The corresponding Pos@Top curve on validation data at different steps.	68
5-1	The overview of the proposed method, robust top-rank learning. . . .	73
5-2	(a) One-dimensional artificial dataset. The blue box indicates the outlier. (b) Results of the original top-rank learning and the robust top-rank learning. The red vertical red bar in (b) indicates the negative outlier location. A digit on a curve indicates the value of p	77
5-3	(a) Two-dimensional artificial dataset. (b) Results of the original top-rank learning and the robust top-rank learning. In the heatmap representation, purple indicates a lower ranking score, and yellow indicates a higher ranking score. The blue box indicates the outlier location. . .	77
5-4	Examples from the Messidor dataset. The first and the second row respectively show the disease images (positive) and the normal images (negative).	78
5-5	t-SNE visualizations of initial feature vectors (by CABNET [1]) of the training (left) and test (right) samples of the Messidor dataset. The red dots denote negative samples, and the green dots denote positive samples. The red dots with the blue box indicates the negative outliers.	79

5-6	t-SNE visualizations results of the comparative methods on the Mes- sidor dataset. Red, green, and purple dots denote negative, positive, absolute top (i.e., doubtlessly positive) samples, respectively; yellow dots with a red circle (pointed by a red arrow) denote the top-ranked negative sample.	80
5-7	(a) Kendall Tau distance and (b) Spearman correlation coefficient. Those values by TopRank CNN and the proposed method are depicted in vertical axes and horizontal axes, respectively.	82

List of Tables

3.1	The mean AUC of RankSVM and Standard SVM on different datasets	47
4.1	<i>Pos@Top</i> and AUC on the test set of the unbalanced CIFAR-10. . . .	59
4.2	<i>Pos@Top</i> and AUC of TopRank CNN (ours) and comparative methods on the DDSM Mammography dataset.	63
5.1	Pos@Top on the Messidor dataset. For a fair comparison, the same network architecture have been used for all the methods.	80

Chapter 1

Introduction

1.1 Background

Learning to rank refers to machine learning techniques used for training ranking models [2], and it aims to organize the objects in a particular order according to their preference, relevance, or other ranking criterion. In the past several decades, learning to rank has attracted the focus of many machine learning researchers because of its growing application in the areas like information retrieval (IR) and recommendation systems [2, 3, 4, 5].

Fig. 1-1 (a) describes the general conception of the bipartite ranking [6], which is one of the most popular ranking methods. If it is possible to separate samples into positive and negative classes, the bipartite ranking methods aim to optimize the ranking function to assign higher ranks to positive samples than all negative samples. As outlined in Fig. 1-1 (a), it depicts a linear ranking function and the samples on a dotted line have the same rank value, such as r_3 . The bipartite ranking approaches have interesting characteristics that are favorable for various applications, such as document retrieval, recommendation systems, and medical image analysis [6, 7, 8, 9], and one important characteristic is that the optimization of bipartite ranking is equivalent to the Area Under the Curve (AUC) maximization [10, 11, 12], which is

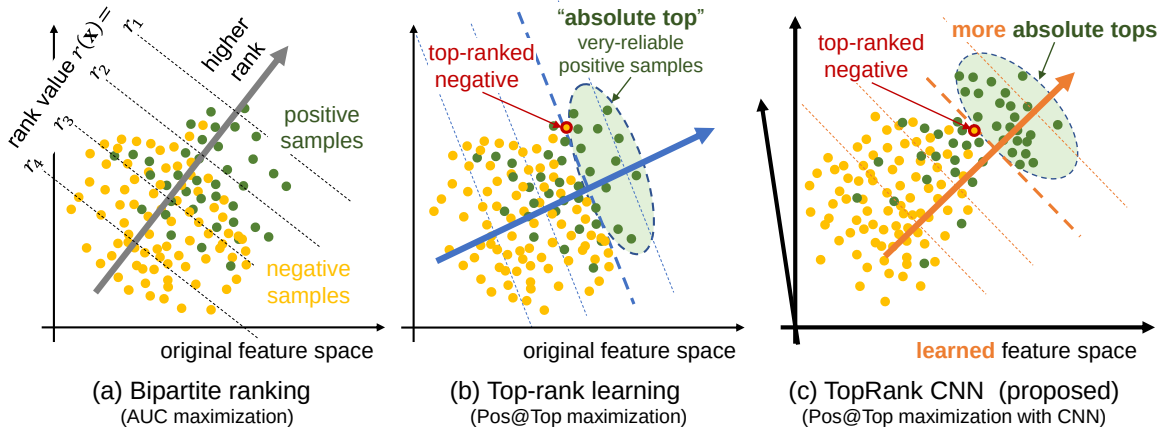


Figure 1-1: Overview of (a) bipartite ranking, (b) top-rank learning, and (c) TopRank CNN (the proposed method). The green dots denote the positive samples and the orange dots denote the negative samples.

very useful for improving the ranking accuracy.

Top-rank learning is a version of the bipartite ranking problem. It emphasizes the ranking accuracy at the top-ranked instances of the rank list, owing to the users who are interested only in those instances [13, 14, 15, 16, 17, 18]. Fig. 1-1 (b) illustrates the concept of top-rank learning. The main difference from bipartite ranking is that top-rank learning attempts to maximize the number of *absolute top* samples instead of maximizing the AUC. *Absolute top* refers to the positive samples ranked higher than the top-ranked (highest-ranked) negative samples. Therefore, *absolute top* samples are *highly reliable positive samples* because they are clearly different from the negative samples. *Pos@Top* is the ratio of the *absolute top* samples to all the positive samples. The property of maximizing Pos@Top will enable the top-rank learning method to be very useful for tasks that require reliable decisions.

1.2 Motivation

1.2.1 Usefulness of Bipartite Ranking

In bipartite ranking problems, samples are labeled positive or negative, and the goal is to learn a ranking function that gives a higher rank value for a positive instance than a negative instance. The bipartite ranking problems have been widely studied in recent years [3, 13, 19, 20]. The bipartite ranking approach has several useful characteristics for improving the unbalanced learning tasks. First, bipartite ranking provides optimal performance for *detecting* positive samples among all samples by pushing positive samples with higher ranking scores by maximizing AUC. Second, the bipartite ranking function is robust to unbalanced data [21]. As it is generally optimized via consideration of all pairs of positive and negative samples, the difference between the numbers of positive and negative samples becomes insignificant. This property implies that the traditional bipartite ranking methods are very useful for solving the data imbalance problem in various applications, such as information retrieval, signature verification, medical image analysis tasks etc [22, 23, 24, 25]. For example, Rakotomamonjy [26] has proved that RankSVM [3] is robust to class imbalance (because of its ability to maximize AUC). Cruz et al. [27] experimentally confirmed this merit. Therefore, as the preliminary research, exploring the properties of the bipartite ranking approach (RankSVM) through the signature verification task becomes one of motivation of this thesis.

1.2.2 Properties of Top-Rank Learning

The objective of top-rank learning is to maximize the Pos@Top, instead of AUC. Considering that the receiver operating characteristic (ROC) of a ranking task is drawn by changing the rank value as the threshold, Pos@Top is observed as the length of the initial (i.e., leftmost) vertical part of the ROC, as shown in Fig. 1-2 (a).

The maximization of Pos@Top is often favorable compared with the maximiza-

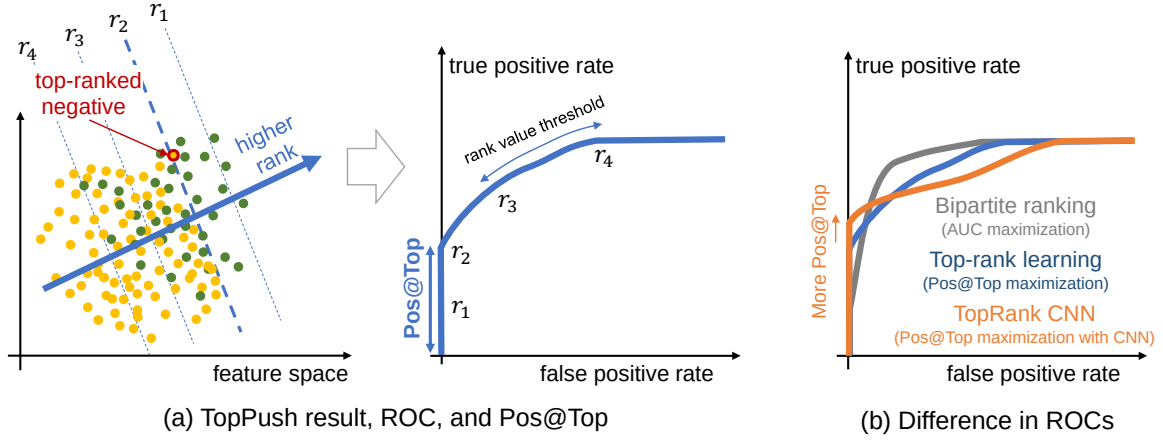


Figure 1-2: Relationship between the rankings and their ROC.

tion of AUC, especially for medical image diagnosis that sometimes relies on medical experts to obtain diagnosis results by manually inspect medical images. Assume that positive and negative samples are images of cancerous and healthy tissues, respectively. The top-rank learning method could learn a model that obtains more absolute tops in the training phrase by optimizing the Pos@Top. If a test image becomes one of *absolute top* samples (that is, its rank value is larger than the top-ranked negative sample), medical experts can skip the manual inspection of the test image; this is because it (clearly) depicts cancerous tissue ¹. In other words, the top-rank learning method could be used as a computer-aided diagnosis technique to provide the most reliable cancerous / clearly normal results automatically, and the medical experts will ignore them in future manual diagnoses. Therefore, if a larger Pos@Top (i.e., more *absolute top* samples) is obtained, manual inspection efforts of medical experts can be reduced drastically. Note that top-rank learning inherits the robustness for unbalanced data. Accordingly, this property provides another merit related to the use of top-rank learning for medical diagnosis.

¹Conversely, it is possible to assume that positive and negative samples denote images of normal and cancerous tissues, respectively. In that case, a test image becomes the *absolute top* is “clearly normal” and the medical experts can also ignore it.

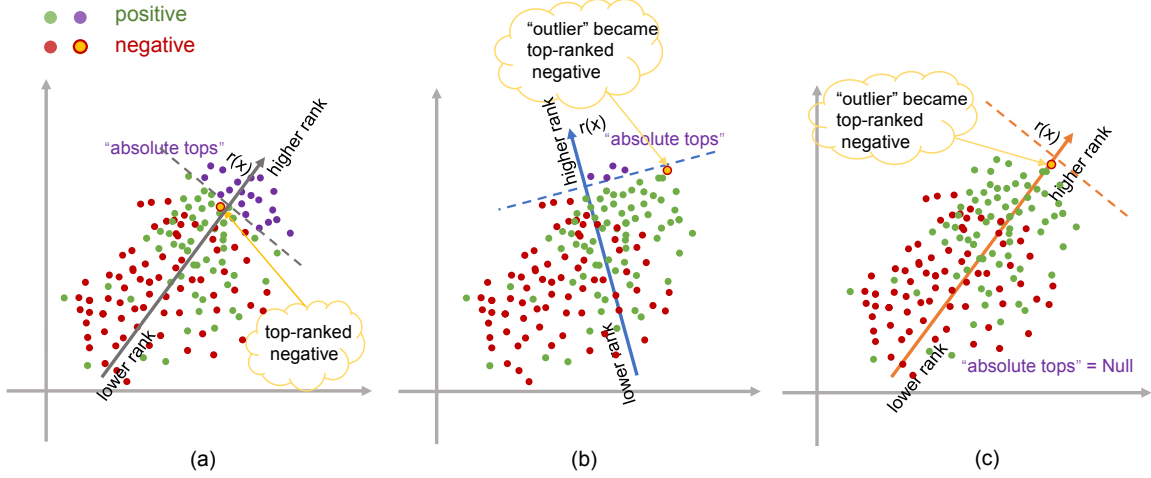


Figure 1-3: (a) Top-rank learning, and (b) its weakness to the negative outlier in the training set. In (b), the negative outlier becomes the top-ranked negative and, consequently, Pos@Top becomes much less than (a). Moreover, the ranking function becomes very different from (a) just by the single outlier. (c) The proposed method can weaken the effect of the outlier and get a more suitable ranking function like (a). Note that this figure does not focus on representation learning (i.e., feature extraction), although the proposed method does it along with top-rank learning in an end-to-end manner.

1.2.3 Limitations in the Conventional Top-Rank Learning

As is discussed in Section 1.2.2, top-rank learning is very useful in various applications that require reliable decisions, but it still has two limitations toward its practical application. Conventional top-rank learning tasks have been formulated in the forms of linear or kernel-based problems that can be solved by various algorithms, such as the support vector machine (SVM)-based algorithm called TopPush [18]. Unfortunately, this formulation cannot fully utilize the favorable properties of top-rank learning. When positive and negative samples cannot be easily separated, it will be difficult to have a large Pos@Top. Although using a kernel function will increase the separability, it is still hard to find an appropriate kernel function for individual tasks. Therefore, how to deal with this limitation of top-rank learning becomes a topic that deserves further researches.

Another serious limitation of top rank learning is the sensitivity to outliers. An

outlier of the negative class will drastically degrade Pos@Top. To observe this, we can compare Fig. 1-3 (a) with (b). Fig. 1-3 (a) presents the top-ranking learning without the effect of outliers. In contrast, 1-3 (b), contains an outlier of the negative class lies in the “very positive” area, As the result, Pos@Top samples become much less than (a). Moreover, just by this single negative outlier, the ranking function in (b) becomes totally different from (a). In short, the ranking function is totally affected by just the single outlier. Unfortunately, in some real applications, especially for medical image analysis tasks, the training set often contains such outliers due to measurement error or labeling error. It becomes necessary to propose a method to weaken the effect of outliers for top-rank learning. A simple solution is to remove outliers from the negative class directly. However, finding an appropriate threshold for outliers is very tough, and removing samples may lose valuable information for the training model. Therefore, one of the motivations of this thesis is to propose a more robust method to weaken the negative effect of outliers.

1.3 Thesis Objectives and Solutions

1.3.1 A Preliminary Research of Top-Rank Learning

Before discussing the top-rank learning, the first objective of this thesis is to explore and verify the characteristics of the learning to rank method, called RankSVM [3], through signature verification tasks. Signature verification systems are formulated as a classification task and suffer from the typical imbalanced problem, where much more negative samples (forgeries) than positive samples (genuine signatures) are available during the training process. Therefore, the signature verification systems impose strict requirements on classifiers. Note that the classifier is optimized to discriminate the genuine signatures and forgery signatures in signature verification systems [28, 29, 30].

RankSVM is one of the learning to rank methods and has been used successfully

to solve the imbalanced learning problems for information retrieval tasks [22, 23, 24]. RankSVM also has several favorable properties for signature verification systems. First, the AUC maximization ability of RankSVM helps to improve the recognition precision for the signature verification system because the optimality of the result is guaranteed. Second, RankSVM provides the rank for each sample. Thereby, positive samples with higher reliability achieve higher ranks. For the signature verification task, this property is very useful because RankSVM assigns ranks to genuine signatures higher based on their reliability. Third, it can handle imbalanced problems without additional parameters. As previously noted, traditional signature verification systems typically encounter imbalanced learning problems.

Considering these favorable properties RankSVM, I applied RankSVM to signature verification tasks. To my best knowledge, RankSVM has not been used for the signature verification task in spite of its usefulness to the task.

1.3.2 Proposal of TopRank CNN

The second objective of this thesis is to address the first limitation of top-rank learning, as discussed in Section 1.2.3. In recent decades, Deep Neural Networks (DNN) has been widely used in image analysis task with great success [31, 32]. The key advantage of DNN lies in the ability to mine representative information from raw data as appropriate feature vectors. For improving the performance of the top-rank learning on complex medical image analysis tasks, this thesis proposes a novel deep neural network *Top-rank convolutional neural network* (TopRank CNN) that enables top-rank learning maximizing $Pos@Top$ with representation learning by DNN. Fig. 1-1 (c) illustrates the ranking function of TopRank CNN. Its feature space is automatically derived by the representation learning from positive and negative samples and therefore different from Fig. 1-1 (a) and (b). If the feature space is optimized for top-rank learning by CNN-based representation learning, we have more reliable *absolute top* samples, i.e., a larger $Pos@Top$. Fig. 1-2 (b) also describes the better performance

of TopRank CNN at the leftmost vertical part of the ROC than the conventional top-rank learning method. These characteristics of TopRank CNN will enhance the usefulness of top-rank learning for medical image diagnosis and various applications that need reliable decisions.

It should be emphasized that combining the ideas of top-rank learning and CNN is not straightforward owing to the following two points. First, using the original objective function of the top-rank learning problem for TopRank CNN easily causes heavy-overfitted results to the training data. This is because the feature representation by CNN tries to obtain the maximum Pos@Top of the training data. Therefore, the p -norm relaxation is introduced into the original objective function to avoid the overfitting problem. The experimental results will prove that the TopRank CNN can achieve larger Pos@Top values under this relaxation with an appropriate hyperparameter, p . Second, the organization of each minibatch requires more consideration than ordinary CNN training tasks. Given that the loss function is calculated by positive-negative pairs (like bipartite ranking), it is essential to ensure that an appropriate number of positive and negative samples are included in each minibatch during the training of TopRank CNN.

1.3.3 Proposal of Robust Top-Rank Learning

The third objective of this thesis is to weaken the effect of outliers for top-rank learning and improve its robustness to outliers. As described in Section 1.2.3, Fig. 1-3 (a) presents the general ranking function of top-rank learning on the training set without negative outliers in negative class, and Fig. 1-3 (b) shows the weakness of the top-rank learning when it suffers from negative outliers.

To solve this problem, a robust top-rank learning algorithm with an unsupervised outlier estimation technique is proposed. The unsupervised outlier estimation technique can estimate the degree of the anomaly of each sample based on the sample density around the sample. (Roughly speaking, if there is no other negative sample

around a negative sample, its degree of anomaly becomes large.) By newly introducing the degree of anomaly into the objective function of the top-rank learning problem, the unexpected effect of the outliers on the ranking function can be weakened. In the following, the standard *local outlier factor* (LOF) [33] is employed as the outlier estimator, although other (even more elaborated) outlier estimators that fit a specific application can be used.

Fig. 1-3 (c) illustrates the expected effect of the proposed method. Although there is the same negative outlier as Fig. 1-3 (b), the proposed method will give a similar ranking function as Fig. 1-3 (a); it is because the proposed method can weaken the effect of the negative outliers during training the ranking function.

This thesis further introduces the representation learning mechanism into the above robust top-rank learning framework, just like TopRank CNN. It can be realized by training a deep neural network (DNN) with the objective function of robust top-rank learning. In other words, it can train the ranking function and the feature extraction function in an end-to-end manner while weakening the effect of outliers.

Experiments are conducted to observe the characteristics of the proposed method on artificial datasets and its performance on practical ranking tasks of medical images. As noted above, the medical image analysis task is a very important target of top-rank learning; however, it often suffers from outliers due to the nature of the labeling difficulty. Therefore, the application of the original top-rank learning to medical images encounters drastic performance degradation. It also shows that the proposed method can weaken the effect of outliers in a medical image set and achieve more stable and superior performance than several comparative methods.

1.4 Novelty and Contributions

The novelty and contributions of this thesis are summarized as follows.

- This thesis firstly utilizes RankSVM for the signature verification task. RankSVM

possesses favorable properties, especially for this task. For example, RankSVM can not only deal with the imbalanced problem without any heuristic remedy but also rank the samples according to their reliability. The effectiveness of RankSVM is verified for signature verification tasks by conducting several extensive experiments. The desired results were achieved on the imbalanced bitmaps of GPDS-150, GPDS-300, GPDS-600, and GPDS-1000 datasets.

- This thesis proposes TopRank CNN, which is a combination of top-rank learning and representation learning for medical image analysis in an end-to-end way. The introduction of p -norm relaxation of the loss function enables the end-to-end training framework of top-rank learning and representation learning which further enhances the usefulness of top-rank learning. Extensive experiments for the proposed TopRank CNN are conducted on medical image-based diagnosis tasks. The results prove the proposed method can achieve more Pos@Top than other methods; this means that the proposed TopRankCNN can find more highly reliable positive samples automatically, and it will reduce the manual inspection efforts of medical experts.
- This thesis attempts to introduce unsupervised outlier estimation into the top-rank learning algorithm for increasing the robustness to outliers. (In fact, it is the first time to assert the importance of robustness against outliers in real-world applications of the top-rank learning algorithm). The usefulness of the robust top-rank learning method via a medical image diagnosis task is also confirmed after confirming its ability to avoid the severe degradation by outliers through experiments with artificial datasets.

1.5 Thesis Structure

This thesis is organized as follows.

- Chapter 1: This chapter gives the general introduction of this thesis. At first, it describes the background of the learning to rank methods and introduces bipartite ranking and top-rank learning. Then, it introduces the motivations, objectives, and solutions on the real-world applications (offline signature verifications and medical image analysis). Finally, it summarizes the novelty and contributions of this thesis.
- Chapter 2: This chapter illustrates more details of learning to rank. First, it introduces the conception of learning to rank. Then, it introduces one of the bipartite ranking methods, RankSVM, and presents its formulation and equations. Finally, it describes the definition and formulation of the top-rank learning methods.
- Chapter 3: This chapter provides detailed information about the utilizing of RankSVM for signature verification tasks. Firstly, it introduces the purpose of using RankSVM for signature verification. Then, it illustrates the related work about different deep learning methods in signature verification tasks and dealing with the data imbalance problem in signature verification tasks. Then, it presents detailed information about the implementation of employing RankSVM for signature verification. Finally, it demonstrates the experimental results of the RankSVM on the signature verification task.
- Chapter 4: This chapter introduces the Top-rank Convolutional Neural Network (TopRank CNN) and its application to medical image analysis. Firstly, it introduces the purpose of the proposal of TopRank CNN. Then, it illustrates the related work of learning to rank with DNN, including the past attempts on top-rank learning with DNN. Then, it presents the definition of the top-rank loss and TopRank CNN. Finally, it presents the experimental results of the TopRank CNN on an unbalanced CIFAR-10 dataset and demonstrates the superiority of the proposed method on medical image analysis.

- Chapter 5: This chapter introduces robust top-rank learning. Firstly, it describes its purpose. Next, it introduces the related work of various outlier detection methods. Then, it presents the definition of the robust top-rank learning method and robust top-rank convolutional neural networks (robust TopRank CNN) and analyzes their advantages. Finally, it shows the experimental results of the proposed robust TopRank CNN on the artificial dataset and demonstrates experimental results with quantitative and qualitative results on medical image analysis.
- Chapter 6: It summarizes the conclusion of this thesis and discusses the future work in this chapter.

Chapter 2

A General Description of Learning to Rank

2.1 Learning to Rank

Techniques for learning to rank have been developed and applied to information retrieval in recent decades [2, 3, 4, 5, 34]. In general, those methods apply machine learning techniques to solve the ranking problem [2]. The learning to rank methods are generally divided into three categories [2, 35]: pointwise, pairwise, and listwise approaches. The pointwise approach, such as subset regression [36], views every single instance as the learning sample. The pairwise approach, such as RankSVM [37] and RankNet [38], treats each pair of instances as the learning sample. The listwise approach, such as ListNet [35] and ListMLE [39], takes the entire ranked list of instances as the learning sample. These methods aim to learn a ranking function to assign rankings to samples by optimizing the corresponding objective function.

This thesis mainly focuses on bipartite ranking [6, 8] which is a special case of pairwise ranking and has received much attention in recent years for binary ranking or classification tasks [7, 8, 12, 40, 41].

In the bipartite ranking problem, samples come from two classes, positive and

negative. The learner is given a positive sample set $\Omega^+ = \{\mathbf{x}_1^+, \dots, \mathbf{x}_m^+\}$ and a negative sample set $\Omega^- = \{\mathbf{x}_1^-, \dots, \mathbf{x}_n^-\}$. The goal of bipartite ranking is to learn a ranking function $r(\mathbf{x})$ that ranks positive samples in Ω^+ higher than negative ones in Ω^- , i.e., that assign higher scores to positive samples than negative ones.

As discussed in Chapter 1, Hanley and McNeil [10] have proved the equivalence between the minimization of the objective of the typical bipartite ranking problem and the maximization of AUC. The AUC is defined as

$$\text{AUC} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(r(\mathbf{x}_i^+) > r(\mathbf{x}_j^-)), \quad (2.1)$$

where $I(\cdot)$ is the indicator function. Consequently, the objective function to be minimized can be defined as follows:

$$\mathcal{J}_{\text{bipartite}} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(r(\mathbf{x}_i^+) \leq r(\mathbf{x}_j^-)). \quad (2.2)$$

2.2 RankSVM

RankSVM [3] is one of the bipartite ranking algorithms when the training data are separated into positive and negative classes, which has various favorable properties as discussed in Section 1.3.1 for the signature verification task. In contrast to the standard SVM, RankSVM aims to learn a real-valued ranking function that gives the positive samples higher values compared to negative samples as much as possible, just as shown in Fig. 2-1. With an imbalanced distribution of the positive and negative classes in Fig. 2-1 (a), the standard SVM will build a hyperplane to classify the genuine samples and forgeries. In Fig. 2-1 (b), the standard SVM does not care about the rank scores of the features. In the worst case, a forgery sample becomes the sample with the largest likelihood as the genuine samples.

Compared to the standard SVM, RankSVM can give the best pattern order by maximizing AUC. For the distribution of Fig. 2-1 (a), RankSVM will give the result

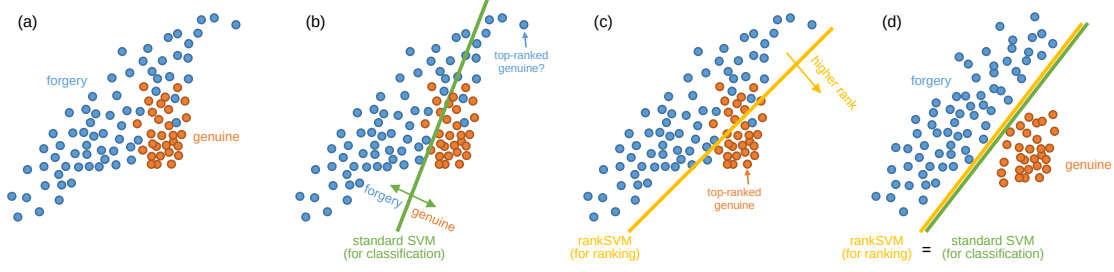


Figure 2-1: Overview of RankSVM. (a) Sample distributions of the positive class (genuine) and the negative class (forgery). The orange dots denote genuinely and the blue dots denote the forgery. (b) Result of standard SVM for classification. In the worst case, a forgery sample becomes the sample with the largest likelihood as the genuine, although it can minimize the classification error. (c) Result of RankSVM, which can give the best pattern order, or equivalently, maximize the AUC. (d) If positive and negative classes are linearly separable, standard SVM and RankSVM give the same result.

like Fig. 2-1 (c), where the most positive samples achieve higher ranks. RankSVM guarantees the generalization performance for imbalanced data because its optimization is equivalent to the AUC maximization. Note that if the original data are linearly separable in some feature spaces, the RankSVM and standard SVM will obtain similar classification boundaries, as shown in Fig. 2-1 (d). Let us start from the optimization problem of the standard SVM. For positive sample set $\Omega^+ = \{\mathbf{x}_1^+, \dots, \mathbf{x}_m^+\}$ and negative sample set $\Omega^- = \{\mathbf{x}_1^-, \dots, \mathbf{x}_n^-\}$, the standard SVM is formulated as follows:

$$\begin{aligned}
 \min_{\mathbf{w}, \xi^+, \xi^-} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i=1}^m \xi_i^+ + \sum_{j=1}^n \xi_j^- \right) \\
 \text{s.t.} \quad & (\mathbf{w}^T \mathbf{x}_j^+ + b \geq 1 - \xi_j^+) \wedge (\xi_j^+ \geq 0), \quad j \in [1, m] \\
 & -(\mathbf{w}^T \mathbf{x}_j^- + b) \geq 1 - \xi_j^- \wedge (\xi_j^- \geq 0), \quad j \in [1, n],
 \end{aligned} \tag{2.3}$$

where ξ_i^+ and ξ_j^- are slack variables corresponding to $\mathbf{x}_i^+$ and $\mathbf{x}_j^-$, respectively. The key idea of RankSVM, which differentiates RankSVM from the standard SVM, is to consider that $\mathbf{w}$ is a ranking function and the rank of a positive sample, i.e., $\mathbf{w}^T \mathbf{x}_i^+$, is better to be larger than that of a negative sample, i.e., $\mathbf{w}^T \mathbf{x}_j^-$. Namely, the number of

pairs which satisfy $\mathbf{w}^T(\mathbf{x}_i^+ - \mathbf{x}_j^-) > 0$ should be maximized. Consequently, this rank optimization is equivalent to AUC maximization because AUC defined in Eq. 2.1 can be rewritten as follows:

$$\text{AUC} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(\mathbf{w}^T \mathbf{x}_i^+ > \mathbf{w}^T \mathbf{x}_j^-), \quad (2.4)$$

where $I(\cdot)$ denotes the indicator function. It should be emphasized that AUC is a common criterion for imbalanced problems (such as information retrieval problems and target detection problems), and thus RankSVM is robust to the imbalance problems.

The optimization problem of RankSVM is formulated as

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \sum_{j=1}^n \xi_{ij} \\ \text{s.t.} \quad & (\mathbf{w}^T(\mathbf{x}_i^+ - \mathbf{x}_j^-) \geq 1 - \xi_{ij}) \wedge (\xi_{ij} \geq 0), \\ & \forall i \in [1, m], \forall j \in [1, n], \end{aligned} \quad (2.5)$$

where ξ_{ij} is a slack variable for the pair of $\mathbf{x}_i^+$ and $\mathbf{x}_j^-$. The first inequalities represent the soft margin constraints so that the rank difference between $\mathbf{x}_i^+$ and $\mathbf{x}_j^-$, i.e., $\mathbf{w}^T(\mathbf{x}_i^+ - \mathbf{x}_j^-)$, becomes positive.

2.3 Top-Rank Learning

Top-rank learning is a special case of the bipartite ranking approach and widely used to optimize the accuracy at the top of a rank list. Interestingly, there are several variations in the formulation of top-rank learning based on the use of different criteria [13, 14, 15, 42, 43, 44, 45]. For example, Joachims [13] proposed a ranking method that maximizes the precision of the top- K ranked samples, often dubbed *precision@K*, which denotes the fraction of positive samples in the top- K ranking.

Given the positive samples set $\Omega^+ = \{\mathbf{x}_1^+, \dots, \mathbf{x}_m^+\}$ and negative samples set $\Omega^- = \{\mathbf{x}_1^-, \dots, \mathbf{x}_n^-\}$, and the label vector $\mathbf{y} \in \{0, 1\}^{m+n}$, the *precision@K* is defined as:

$$\text{Prec@K} = \frac{1}{K} \sum_{s=1}^{m+n} y_s I(r(\mathbf{x}_s) - r(\mathbf{x})_{[K]}), \quad (2.6)$$

where s is the index of samples in $\Omega^+ + \Omega^-$ and $y_s = 1, 0$ denote the label of positive sample and negative sample. $I(\cdot)$ is an indicator function, K is a small constant, usually set to 3, 5, 10. The $r(\mathbf{x})_{[K]}$ denotes the ranking score the k -th largest sample among all samples. The minimization of objective function is equivalent to the maximization of *precision@K*. Consequently, the objective function can be written as:

$$\mathcal{J}_{\text{Prec@K}} = \sum_{s=1}^{m+n} (1 - y_s) I(r(\mathbf{x}_s) - r(\mathbf{x})_{[K]}). \quad (2.7)$$

In [15], another formulation *Average Precision (AP)* that focused on the top-ranked samples was found. The AP is calculated by the average of *precision@K*:

$$\mathcal{J}_{\text{AP}} = \frac{1}{m} \sum_{K=1}^m (\text{Prec@K}). \quad (2.8)$$

Narasimhan and Agarwal [14] maximized partial AUC, which evaluates the AUC between a certain ROC interval. Its formulation is defined as:

$$\mathcal{J}_{\text{pAUC}} = \frac{1}{mn(\beta - \alpha)} \sum_{i=1}^m \sum_{j=1}^n I(r(\mathbf{x}_i^+) - r(\mathbf{x}_j^-)), \quad (2.9)$$

where α and β denote the specific false positive rates, and $0 \leq \alpha < \beta \leq 1$. Despite the theoretical interests of these formulations, they result in non-convex optimization problems that cannot be solved efficiently.

According to the variations of the formulations, the algorithms for top-rank learning also exhibit other variations. Specifically, optimizing the *Pos@Top* became popular in recent literatures [18, 46, 47, 48].

As shown in Figs. 1-2 (a) and (b), Pos@Top is the ratio of absolute top samples among all m positive samples. The absolute top samples are the “most reliable positive” (or “very positive”) samples, so maximizing Pos@Top is beneficial to many applications. Formally, Pos@Top is defined as

$$\text{Pos@Top} = \frac{1}{m} \sum_{i=1}^m I(r(\mathbf{x}_i^+) > \max_{1 \leq j \leq n} r(\mathbf{x}_j^-)), \quad (2.10)$$

where $\max_{1 \leq j \leq n} r(\mathbf{x}_j^-)$ denotes the ranking score of top-ranked negative samples. The indicator function $I(\cdot)$ is replaced by a surrogate function $\ell(\cdot)$, which is a convex, non-decreasing, and differentiable function. Consequently, the loss function to be minimized is formulated as follows.

$$\begin{aligned} \mathcal{J}_{\text{Pos@Top}} &= \frac{1}{m} \sum_{i=1}^m \ell \left(\max_{1 \leq j \leq n} r(\mathbf{x}_j^-) - r(\mathbf{x}_i^+) \right) \\ &= \frac{1}{m} \sum_{i=1}^m \max_{1 \leq j \leq n} \ell \left(r(\mathbf{x}_j^-) - r(\mathbf{x}_i^+) \right). \end{aligned} \quad (2.11)$$

The negative sample $\mathbf{x}_j^-$ that gives the maximum is the top-ranked negative.

By optimizing the Pos@Top, the top-rank learning method can produce more reliable results, which is very useful for tasks that require reliable decisions. In later experiments, top-rank learning will be employed to optimize the neural networks and apply them for medical image diagnosis tasks.

Chapter 3

RankSVM for Offline Signature Verification

3.1 Purpose

Signature verification systems aim to verify a person’s identity by recognizing their handwritten signature, which is crucial for enhancing security and privacy in various fields, including banks, educational institutes, attendance verification, and official document verification, where authentication is of utmost importance. Signature verification systems suffer from imbalanced learning problems when it learns the classifier to discriminate the genuine and forgeries. The standard classification approaches, such as SVM, often degrades the performance for imbalanced data or requires additional parameters for data balancing.

To solve this problem, RankSVM is proposed for signature verification tasks. RankSVM is a bipartite ranking method, which theoretically guarantees the generalization performance for imbalanced data. In addition, RankSVM can rank samples and thereby give the most reliable genuine signatures higher rankings among all samples. To investigate the ability of RankSVM in signature verification tasks, extensive experiments is conducted on bitmaps of GPDS-150, GPDS-300, GPDS-600,

and GPDS-1000 datasets and deep features of the GPDS-960 dataset.

The following first reviews the related work and then discusses the implementation of RankSVM for signature verification tasks. Finally, it illustrates the experimental results and summarizes the usefulness of RankSVM.

3.2 Related Work

3.2.1 Deep Learning methods for Signature Verification

Signature verification can be considered a two-class classification problem: to judge whether two signatures are signed by the same person or to classify whether a signature is genuine for a specific user. In recent years, many deep learning-based approaches have been proposed for the signature verification task [30, 49, 50, 51, 52, 53, 54]. According to the degree of simulation, signature forgeries are categorized into random forgery, simple forgery, and skilled forgery as [55, 56]:

- Random forgery: genuine signature of any writer other than the authentic writer.
- Simple Forgery: a forger only knows the name of the authentic author.
- Skilled Forgery: the forger has access to authentic signatures and forges after practicing an unrestricted number of times.

Signature verification is generally divided into two categories: online and offline signature verification according to the different devices of signature data acquisition. In offline signature verification, the signature is captured by a scanner or camera after the writing process is completed and represented as a digital image. Since the dynamic information of stokes (e.g., pressure, the altitude of the pen) is lost, offline signature verification has become a very challenging problem [57]. The general offline signature verification system usually consists of three steps of preprocessing,

feature extraction, and feature verification [54, 56, 58]. The preprocessing operations mainly include background elimination, noise reduction, signature normalization, and so on [56, 59]. Feature extraction is one of the important parts of the signature verification system. To obtain good representations for signatures, many kinds of literature proposed various deep learning-based methods to learn good representations for capturing the behaviors of different writers or discriminate genuine against forgery [30, 49, 50, 54]. For instance, Hafemann et al. [30, 50] proposed a CNN-based feature extraction approach, named "Signet," to obtain discriminative features between genuine signatures and skilled forgeries for different users. Based on this work, Zheng et al. proposed to capture the discriminant information not only between the genuine signatures and skilled forgeries but also different users by extracting the micro deformations in the max-pooling operation, which achieved the better performance than the state-of-the-art method [54].

Feature verification is the final step of signature verification, which aims to train a writer dependent [52, 60] or writer independent [61] classifier for discriminating the genuine or forgery [30]. The most common approach is to design Writer-dependent classifiers. In this scenario, a training set is constructed for each user of the system, consisting of genuine signatures as positive examples and random forgeries as negative samples. A binary classifier is then trained on this dataset, resulting in one model for each user. Compared to the writer-independent approach that trains one model for all users, the writer-dependent approach, in general, obtains higher accuracy [30]. This work proposes to train a writer-dependent classifier to distinguish the genuine and skilled forgery signatures.

3.2.2 Imbalanced Problem in Signature Verification

In the signature verification task, the imbalanced learning problem often occurs in the training process, making it difficult to build robust verification systems, especially for training the classifier in feature verification phrases. Hafemann et al. [29, 30, 50]

used a version of SVM which cares the class-imbalance [62]. However, this SVM requires additional parameters for different classes, and thus it has to tune with more computational costs. Note that additional parameters quadratically increase the computational cost when using grid search. Although there is a simple heuristic to decide the additional parameters, there is no theoretical justification.

To solve the imbalanced learning problem in traditional SVM, Joachims [3] originally proposed RankSVM and applied it to web search engines. RankSVM solves AUC maximization based on the formulation of standard SVM (the details are shown in Section 2.2) without any additional parameters to SVM. RankSVM is justified by the theoretical results in the machine learning field (e.g., [63]), which says that RankSVM achieves a high AUC score for test samples (i.e., unseen data) as training samples. Because of the strong theoretical support, RankSVM has been applied to various domains such as document retrieval [3, 22, 23, 24] and image classification [64, 65]. The theoretical and empirical success of RankSVM motivates us to apply it to the signature verification task. To the best of our knowledge, RankSVM has never been used in the signature verification task.

3.3 Utilizing the RankSVM for Offline Signature Verification

3.3.1 Implementation of RankSVM for Skilled Forgery Attacks

RankSVM is applied to the signature verification for skilled forgery attacks, which is a typical imbalanced problem, and compared the performance of RankSVM to the standard SVM. Its objective is to learn a model that can distinguish between genuine signatures and skilled forgeries. In the experiments, a training set for each specific user was built comprising genuine signatures as positive samples and random forgery

(registered genuine signatures from third-persons) as negative samples. As the test samples, skilled forgeries are used, which are signatures for imitating the genuine signatures by other persons.

To enable fair comparison with RankSVM, this thesis uses an implementation of SVM that cares the class-imbalance by weighting different costs for the misclassification of different classes [62]. In this implementation, the standard SVM of Eq. (2.3) is modified as follows:

$$\begin{aligned}
& \min_{\mathbf{w}, \xi^+, \xi^-} \frac{1}{2} \|\mathbf{w}\|^2 + C^+ \sum_{i=1}^m \xi_i^+ + C^- \sum_{j=1}^n \xi_j^- \\
& \text{s.t.} \quad (\mathbf{w}^T \mathbf{x}_i^+ + b \geq 1 - \xi_i^+) \wedge (\xi_i^+ \geq 0), \quad i \in [1, m] \\
& \quad \quad (-(\mathbf{w}^T \mathbf{x}_j^- + b) \geq 1 - \xi_j^-) \wedge (\xi_j^- \geq 0), \quad j \in [1, n].
\end{aligned} \tag{3.1}$$

Hyper-parameters are set as $C^+ = \frac{n}{m} C^-$ and $C^- = 1$ according to the state-of-the-art signature verification methods [29, 30, 50]. For RankSVM, it simply sets $C = 1$.

3.3.2 Implementation of RankSVM for Signature Retrieval

The behavior of the RankSVM (Eq. 3.1) and standard SVM (Eq. 2.3) are also observed through a signature *retrieval* task, where the largely-imbalanced data is needed to be solved. I used the same training set as in the verification task and built a new testing set consisting of 12 genuine signatures from the user as positive samples and 12 random forgeries as negative samples. Additionally, I used the same hyperparameter settings of the RankSVM and standard SVM as the verification task in the later experiments.

3.4 Experiments for Skilled Forgery Attacks

The experiments were conducted by using the offline signature dataset called “GPDS synthetic online offline signature database” [66]. It contains data from 10,000 syn-



Figure 3-1: Signature examples from GPDS dataset.

thetic individuals: 24 genuine signatures for each individual and 30 skilled forgeries for their signatures. Fig. 3-1 shows the samples of genuine signatures and their corresponding skilled forgeries. In the experiments, 64×64 pixels of bitmaps were used as inputs while the pixel values were normalized to $[0,1]$. Aside from this normalization step, no other preprocessing was applied.

3.4.1 Evaluation Metrics

In signature verification, False Rejection Rate (FRR), False Acceptance Rate (FAR), Equal Error Rate (EER) and Detection Error Tradeoff (DET) curve are used to evaluate the performance of RankSVM and the standard SVM. FRR is the fraction of genuine signatures rejected as forgeries, FAR is the fraction of forgeries accepted as genuine signatures, EER is the error when $FRR = FAR$, and the DET curve is a graphical plot of the error rates for binary classification tasks, which plots FRR versus FAR. A lower EER implies a better performance for the algorithm.

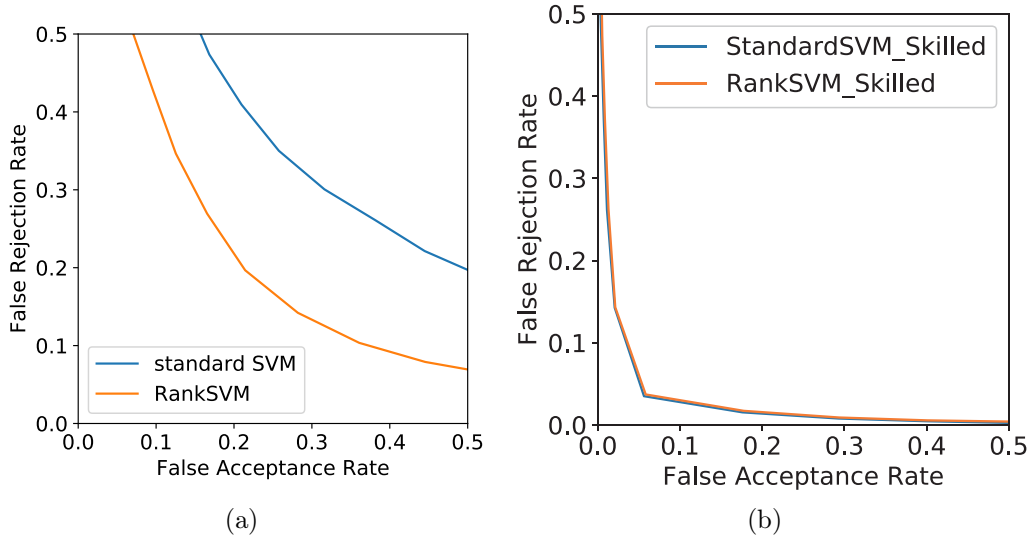


Figure 3-2: (a) The DET curve of RankSVM and the standard SVM with bitmap on GPDS-150 dataset. (b) The DET curve of RankSVM and the standard SVM with deep features on GPDS-960 dataset.

3.4.2 Experimental Result with Bitmap

In this experiment, for each user, I built a training set, which consisting of 12 genuine signatures from the user as positive samples and 12 random forgeries (genuine signatures from other users) as negative samples and a testing set consisting of 10 randomly selected genuine signatures from an user as positive samples and 10 randomly selected skilled forgeries from other users as negative samples with respect to GPDS-150. I created 10 random partitions for training set and testing set and repeated the same experiments for statistical evaluation.

The average EER and standard deviation obtained by RankSVM and the standard SVM were 19.56 ± 0.8620 and 30.14 ± 0.8142 , respectively. The DET curve for a random partition of users is shown in Fig. 3-2 (a). From this curve, it easy to see that the RankSVM performs significantly better than the standard SVM on signature verification task, and the EER of RankSVM is obviously lower than that of the standard SVM.

3.4.3 Experimental Results with Deep Feature

The experiment is also conducted that used the Convolutional Neural Network (CNN) based deep features with the RankSVM and the standard SVM [30]. The deep features on GPDS-960 [30] were directly used to train and test the RankSVM and standard SVM models. The training and testing processes are same as in [30]. For first 300 users of GPDS-960, RankSVM obtained a nearly similar performance with standard SVM with an EER of 4.1% for RankSVM and 3.9% for standard SVM. The DET curve are shown in Fig. 3-2 (b). It needs to consider the reason that a significant difference between RankSVM and the standard SVM is not found when using deep features. A possible reason is that our RankSVM obtained highly similar hyperplane as the standard SVM as shown in the case of Fig. 2-1 (d). In other words, if the hyperplanes of RankSVM and SVM are highly similar, the deep feature space is possibly linearly separable.

To ensure the above reason, the cosine similarity was calculated¹ between the hyperplanes by RankSVM and the standard SVM on the bitmap space (for GPDS-150 dataset), and also calculated the cosine similarity between the hyperplanes by RankSVM and the standard SVM on the deep feature space (for GPDS-960 dataset). It could see the significant difference between the former average cosine similarity ($= 0.776$) and the latter ($= 0.866$). This result implies that the deep feature space gives similar hyperplanes for RankSVM and the standard SVM and thus is much more easily linear-separable than the bitmap space.

Consequently, the suggestion by these observations is as follows: RankSVM is more suitable for tough tasks where the distributions of the positive and the negative samples are heavily overlapping, rather than for easier tasks where these samples are already almost separable by a simple classifier like the standard SVM.

¹The cosine similarity is used because the bias of the hyperplane of SVM can be ignored in ranking tasks.

Table 3.1: The mean AUC of RankSVM and Standard SVM on different datasets

Dataset	RankSVM	standard SVM
GPDS-150	0.970	0.930
GPDS-300	0.965	0.911
GPDS-600	0.955	0.895
GPDS-1000	0.952	0.886

3.5 Experiments in Signature Retrieval

3.5.1 Quantitative Evaluation and Qualitative Evaluation

The performance of RankSVM and the standard SVM is quantitatively and qualitatively evaluated on the signature retrieval task by using AUC and observing the retrieval results. The experiments were conducted by using the datasets of GPDS-150, GPDS-300, GPDS-600, and GPDS-1000 consisting of the first 150, 300, 600 and 1,000 users, so that the imbalance between the positive and the negative samples is increased with the number. The signature retrieval experiments with RankSVM and the standard SVM were performed multiple times by using randomly selected different training and test datasets to obtain statistical performance.

Quantitative Evaluation:

The mean AUC of each dataset for 10 random partitions are presented in Table 3.1, and Fig. 3-3 shows the average ROC curves created for users among these datasets. Table 3.1 and Fig. 3-3 indicate that RankSVM significantly outperforms the standard SVM in the retrieval task. Although the dataset becomes more imbalanced with the increasing number of users, the performance degradation of RankSVM is much less than the standard SVM. This is because the standard SVM is not stable on imbalanced data, whereas RankSVM keeps high AUC performance even for highly imbalanced datasets.

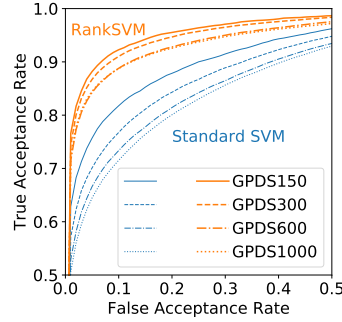


Figure 3-3: The ROC curve of RankSVM and standard SVM on GPDS-150, GPDS-300, GPDS-600, GPDS-1000 datasets.

Qualitative Evaluation

Fig. 3-4 shows the top-5 retrieval results on GPDS-150 dataset. From this figure, the following facts can be confirmed:

1. RankSVM could get genuine samples as the top-ranked samples whereas the standard SVM often fails to get them.
2. Especially, it should be emphasized that the top-ranked samples are very similar to the query sample. This indicates that RankSVM could give the most reliable genuine samples among all genuine samples.
3. It also should be emphasized that the standard SVM ranks a very different signature at the top. This is caused by the situation of Fig. 2-1 (b). In contrast, RankSVM gives more similar signatures to the input.

3.6 Summary

The signature verification task is a typical imbalanced problem, and thus classifiers for the task are better to be data imbalance. The past trials with the standard classification approaches, such as SVM, often employ a simple and heuristic remedy to balance the problem by using weight parameters. In this thesis, RankSVM was

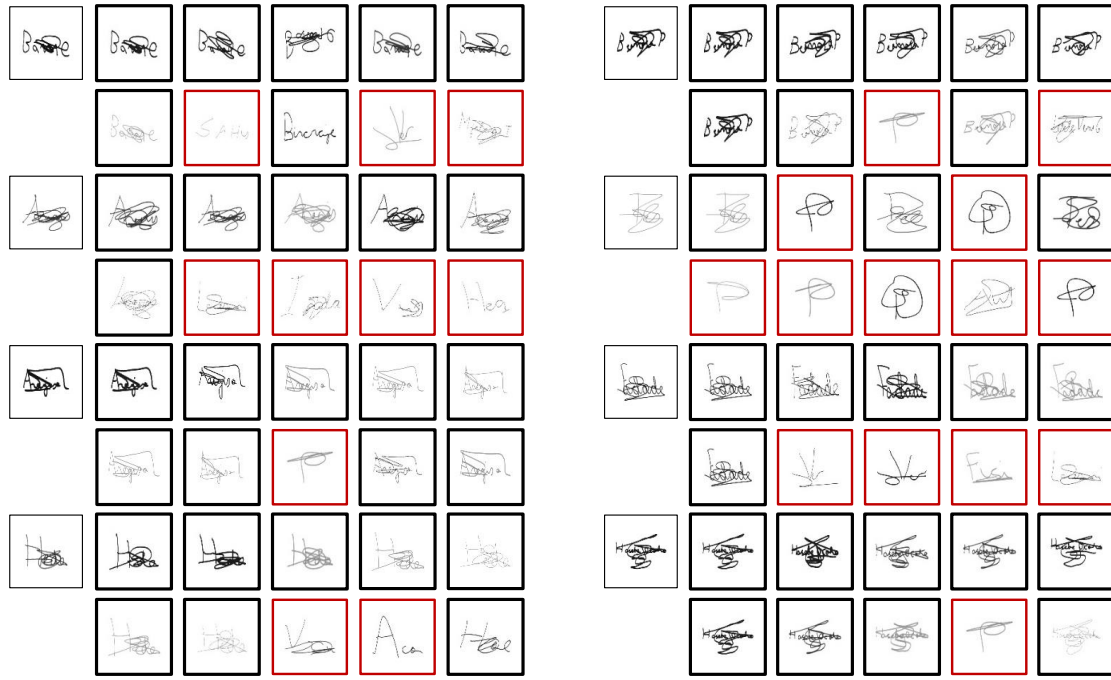


Figure 3-4: Top-5 signatures for RankSVM and Standard SVM on GPDS-150 dataset. Each two rows denote one user's ranked signatures; the top and bottom rows show the top-5 results of RankSVM and the standard SVM, respectively. Images in the black box and the red box denote the correct genuine signatures and the wrong signatures, respectively.

applied in the signature verification task. RankSVM has several good properties for the task. Especially, it theoretically guarantees the robustness to the imbalanced task without any parameters.

The extensive experiments on signature verification and retrieval tasks confirmed that RankSVM could achieve significantly higher performance than the standard SVM for both signature retrieval and signature verification tasks, especially in a tough situation where the positive and the negative classes are heavily overlapping.

Chapter 4

TopRank CNN for Medical Image Analysis

4.1 Purpose

Top-rank learning aims to maximize the number of absolute top samples, which are “doubtlessly positive” samples and beneficial for real applications that require reliable positive samples, such as medical image analysis. However, the conventional top-rank learning tasks are limited in dealing with complicated tasks due to the lack of a powerful representation ability. To solve this limitation of top-rank learning, this chapter proposes a Top-rank convolutional neural network (TopRank CNN) to realize top-rank learning with representation learning for complicated tasks. TopRank CNN inherits the advantages of top-rank learning and representation learning that avoids data imbalance problems and learns a feature space for providing more reliable positive samples.

First, I summarize the related work of learning to rank with deep neural networks and top-rank learning with deep neural networks. Then, I present the detailed explanations for the proposed TopRank CNN. Since the overfitting problem in the original objective function of top-rank learning, the p -norm relaxation is employed to the

original loss function in TopRank CNN. Finally, I prove the usefulness of TopRank CNN experimentally with medical diagnosis tasks that require reliable decisions and robustness to the unbalanced learning problem.

4.2 Related Work

4.2.1 Learning to Rank with Deep Neural Networks

With the development of deep learning, recent studies have applied the idea of learning to rank to a variety of applications, such as information retrieval [67, 68, 69], face recognition [70, 71], and person re-identification [72, 73]. The typical motivation related to combining deep learning and learning to rank is to realize a metric learning method, whereby the similarity is evaluated by a ranking function. For example, Huang et al. [67] proposed the deep structure semantic model (DSSM), and Shen et al. [74] proposed the convolutional deep structured semantic Model (CDSSM) to directly apply deep neural networks to obtain the semantic representations of queries and documents. The ranking score was produced by computing their cosine similarity and achieved good performance for web searching. More recently, Pang et al. [75] proposed DeepRank capture important semantic features for information retrieval. The relevance scores of DeepRank are obtained by using the pairwise ranking loss. DeepRank achieved higher state-of-the-art results compared with traditional learning to rank methods used for information retrieval. In [73], Chen et al. also proposed an effective deep ranking framework combining learning to rank algorithm with deep CNN for solving person re-identification tasks. They aimed to obtain a similarity metric in which the top-k closest candidate images to the query are matched with the query. To obtain accurate results at the top, they applied learning-to-rank techniques with metric learning and CNN.

In recent years, some studies also propose to combine deep learning with learning to rank methods for medical image analysis tasks [76, 77]. For example, Lyu et al. [77]

proposed combining the RankNet with convolutional neural networks for dealing with the ultrasound image selection task, which was first regarded as the ranking problem and achieved better performance than the experience of a human expert.

In these approaches, CNN-based representation learning played an important role in enhancing the effectiveness of traditional learning to rank methods. Furthermore, it offered a feature space wherein the positive samples were pushed to be ranked before any negative samples by maximizing the AUC.

4.2.2 Top-Rank Learning with Deep Neural Networks

Considering the effectiveness of the top-rank learning method on optimizing the accuracy at the top, some studies combined the idea of top-rank learning (by optimizing different criteria, e.g., the precision at the top, the number of absolute tops) with deep learning models [78, 79].

In [78], Song et al. employed top-rank learning to solve the binary coding problem for image data. The target of this work was to learn coding functions that the data in the same group were close to each other in the coding space (Hamming space), which focused on the closeness of the top position (i.e., representative data in the same group) in the Hamming space, and optimized the precision of the top position.

To learn non-linear representation for ranking, Geng et al. used the concept of Pos@Top to obtain discriminative local features in data [79], which achieved better performance on Pos@Top. However, this method just used the Pos@Top-based loss to learn CNN-based representation, not for the classifier. More importantly, this method was designed to solve the multiple-instance problem and hard to solve vectorized data that cannot be split into a set of instances.

This study broke through the limitations of [79] and proposed TopRank CNN to integrate top-rank learning and representation learning using p -norm relaxation by an end-to-end framework. TopRank CNN supported the mapping of samples to the rankable deep feature space and maximized the number of positive samples ranked

before the top-ranked negative samples. In this work, the application of TopRank CNN on medical image diagnosis produced a better performance than the comparative methods.

4.3 Top-Rank Convolutional Neural Networks

This section demonstrates the problem of the combination of top-rank learning and convolutional neural networks and introduces p -norm into the objective function of top-rank learning. After that, the detailed information for the implementation of TopRank CNN is presented.

4.3.1 p -norm Relaxation

Since the objective of the TopRank CNN is to maximize Pos@Top, Eq. (2.11) can be used for the loss function of CNN. However, there is a large risk of overfitting to the training data. CNN has a powerful representation learning ability so as to minimize the loss function [31, 32]. In an extreme case with Eq. (2.11), all the positive samples become the absolute top samples (by making them very different from the negative samples) and thus achieves $Pos@Top=100\%$. Considering the presence of outliers in the training samples, this “perfect” situation does not often work in the case of the test samples.

Therefore, this study introduces the p -norm relaxation of the max operation in Eq. (2.11). Specifically, the following loss function is used, called TopRank loss,

$$\mathcal{J}_{\text{TopRankCNN}} = \frac{1}{m} \sum_{i=1}^m \left(\sum_{j=1}^n (\ell(r(\mathbf{x}_i^+) - r(\mathbf{x}_j^-)))^p \right)^{\frac{1}{p}}. \quad (4.1)$$

If the hyper-parameter $p \rightarrow \infty$, $\mathcal{J}_{\text{TopRankCNN}} \rightarrow \mathcal{J}_{\text{Pos@Top}}$. If p is set at a smaller value, say 16 or 32, our loss function $\mathcal{J}_{\text{TopRankCNN}}$ has a milder effect on the maximization of Pos@Top, and can avoid the overfitting issue. This effect will be proved in a

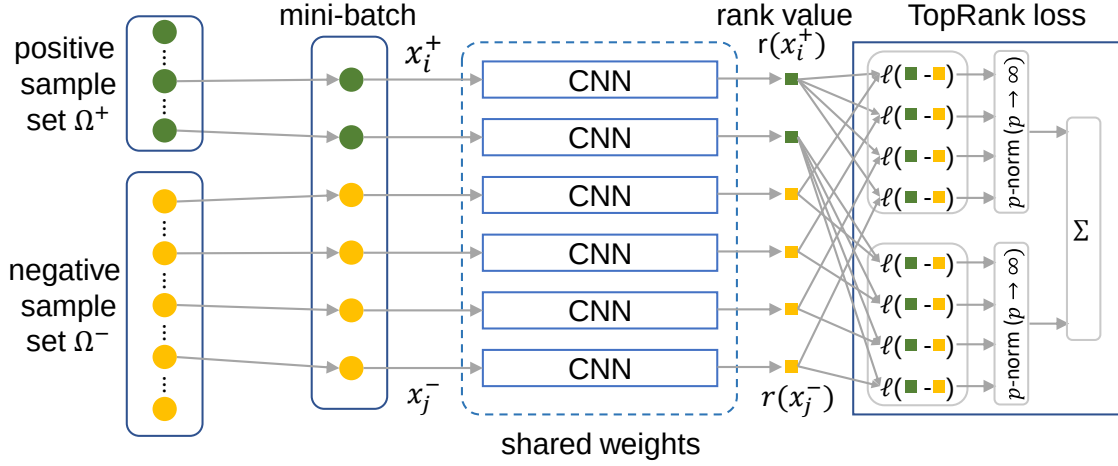


Figure 4-1: The architecture of TopRank CNN.

subsequent experiment. In the later experiments, $\ell(z) = \log(1 + e^{-z})$ is used as the surrogate function. It should be emphasized again that the minimization of Eq. (2.11) does not maximize AUC but $Pos@Top$. Therefore, the ranking function by Eq. (2.11) may give a smaller AUC than Eq. (2.2).

4.3.2 TopRank CNN

Fig. 4-1 shows the network architecture of the proposed TopRank CNN that is designed to integrate top-rank learning and representation learning using an end-to-end framework with the TopRank loss (Eq. 4.1). Based on this architecture, it will have an appropriate feature representation for higher $Pos@Top$ (i.e., more absolute tops), as shown in Fig. 1-1(c).

Given that the TopRank loss function needs positive and negative pairs, Siamese architecture was used [80], which consists of two identical CNNs with the same weights, and used two sample inputs. It should be noted that our TopRank CNN is still very different from typical Siamese networks given its purpose to derive the ranking function $r(\mathbf{x})$ with the TopRank loss (Eq. 4.1) for the maximization of $Pos@Top$.

One technical topic that is unique to TopRank CNN is the organization of samples

in each minibatch. For each minibatch, the TopRank loss (Eq. 4.1) is calculated by using all the pairs of positive and negative samples. Thus, each minibatch needs to contain the samples from both classes (if a minibatch is comprised of the samples from a single class, it is impossible to calculate the loss). More importantly, (Eq. 4.1) is the p -norm relaxation of (Eq. 2.11), and thus its minimization will push the positive samples to have higher ranks than *the top-ranked negative samples within the minibatch*. This means that it is implicitly expected that each minibatch includes a negative sample that is close to the top-ranked negative samples. Consequently, each minibatch is better organized by using more negative samples than positive samples to satisfy the expectation as much as possible.

The value of the hyper-parameter p will be set by using a validation set. In general, a very large p should be avoided, although it makes (Eq. 4.1) similar to its original form (Eq. 2.11). This is because large p values easily cause numerical overflows (scaling techniques for normalizing the input data by dividing its maximum value may avoid overflows with larger p values, but it may result in underflow). Thus, in practice, the validation steps are sometimes needed to be stopped before a larger p value causes an overflow.

4.4 Experiment on an Unbalanced CIFAR-10 Dataset

In the experiments, the comparative models and the proposed TopRank CNN model were implemented using the platform of TensorFlow and the Python 3.5.2 packages. The training and test model was based on the GPU environment of two NVIDIA GeForce GTX 1080 Ti GPU with 10GB RAM.

4.4.1 Dataset

To investigate the basic performance of TopRank CNN for solving unbalanced problems, it is first applied to an unbalanced version of the CIFAR-10 dataset [81]. In the

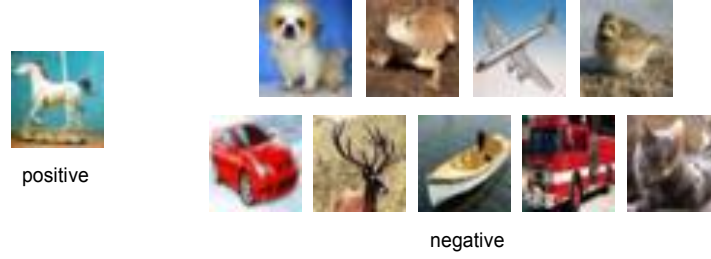


Figure 4-2: Examples of positive and negative classes from the unbalanced CIFAR-10 dataset.

unbalanced CIFAR-10, the class “horses” was selected as the positive class, and the other nine classes were negative class, as shown in Fig. 4-2 (note that the selection of “horses” was random). From the training set of CIFAR-10, 1000 positive and 9000 negative samples are randomly selected to build the validation set. Thus, the training, validation, and testing sets contain 4000, 1000, and 1000 positive samples and 36000, 9000, and 9000 negative samples, respectively. Each sample corresponds to a $32 \times 32 \times 3$ color image.

4.4.2 Comparative Methods

The following experiments were conducted with five comparative methods.

- $\text{TopPush}_{\text{CNN}}$: The original TopPush algorithm [18] used the feature extracted from the previous fully connected layer before the final layer of the following CNN_{soft} .
- $\text{RankSVM}_{\text{CNN}}$: The RankSVM algorithm [3] also used the feature extracted from the previous fully connected layer before the final layer of the following CNN_{soft} .
- CNN_{soft} : CNN trained with the standard softmax loss for binary classification (positive and negative classes).

- $\text{CNN}_{\text{focal}}$: CNN was trained with a focal loss [82] and is one of the most extensively known solutions used to cope with the data-imbalance binary classification problem.
- $\text{CNN}_{\text{pos@top}}$: CNN trained with $\mathcal{J}_{\text{Pos@Top}}$ of Eq. (2.11), instead of its p -norm relaxation.

Note that the same architecture are used for all the CNN-based methods except for the final loss layer. Among all comparison methods, $\text{RankSVM}_{\text{CNN}}$ and $\text{TopPush}_{\text{CNN}}$ used the fixed feature representation given by CNN_{soft} , which aims to compare the performance of the learning to rank method and top-rank learning method on AUC and Pos@Top. By the way, RankNet [4], as another conventional learning to rank method, needs an appropriate feature extraction as its input; in other words, its performance will largely depend on the feature extraction method. To make the comparison as fair and clear as possible, I do not employ RankNet in the comparative study. The comparison between $\text{RankSVM}_{\text{CNN}}$ and the proposed method will establish the superiority of TopPush loss of Eq.(2.11) over the standard learning to rank loss of Eq.(2.2) employed in RankNet.

4.4.3 TopRank CNN Architecture and Hyper-parameters

As shown in Fig. 4-1, TopRank CNN is composed of two identical CNNs (for a positive-negative samples pair of input) that share the same weights. Each CNN consists of three convolutional layers and two fully connected layers, namely, the rectified linear unit (ReLU) activation function and the max-pooling layers that followed each convolutional layer. The learning rate decay method was employed to adjust the learning rate during the training, with an initial learning rate of 0.01, and was updated with a decay factor of 0.90 after every epoch. Each minibatch contained 5 positive and 450 negative samples — this unbalance minibatch is mandatory owing to the reason outlined in Section 4.3.2. The stochastic gradient descent (SGD) algorithm

Table 4.1: *Pos@Top* and AUC on the test set of the unbalanced CIFAR-10.

Method	Feature	Algorithm	Loss	Hyper-param.	AUC Test	Pos@Top	
						Train	Test
TopPush _{CNN}	CNN _{soft}	TopPush	Eq.(2.11)	$\lambda = 10$	0.7186	0.0383	0.0270
RankSVM _{CNN}	CNN _{soft}	RankSVM	Eq.(2.2)	$C = 0.01$	0.8912	0.0475	0.0210
CNN _{soft}	end-to-end	CNN	Softmax	-	0.8804	0.0410	0.0260
CNN _{focal}	end-to-end	CNN	Focal	-	0.9124	0.0318	0.0590
CNN _{pos@top}	end-to-end	CNN	Eq.(2.11)	-	0.8894	0.4635	0.0990
TopRank CNN (ours)	end-to-end	CNN	Eq.(4.1)	$p = 16$	0.9465	0.4910	0.1950

was used to optimize the network.

The validation set was used to determine the hyper-parameter of $p \in \{2, 4, 8, 16, 32, 64\}$ for TopRank CNN, $\lambda \in \{10^{-2}, 10^{-1}, 1, 10^1, 10^2\}$ for TopPush_{CNN}, and $C \in \{10^{-2}, 10^{-1}, 1, 10^1, 10^2\}$ for RankSVM_{CNN}. Note that a larger p causes numerical overflow. Thus, p cannot be larger than 64. The hyper-parameter of class weight $c = m/n$ was also utilized to balance the positive and negative classes for the training of conventional CNN.

4.4.4 Results on the Unbalanced CIFAR-10 Dataset

Table 4.1 shows the AUC and Pos@Top of the proposed method and the five comparative methods. In the cases of the classification methods, such as CNN_{soft}, CNN_{focal}, their soft-max values were used to calculate AUC and Pos@Top. This table indicates the following facts:

- 1) Among all the methods, the proposed TopRank CNN achieved the best performance on Pos@Top, as expected
- 2) TopPush_{CNN} achieved higher Pos@Top, but lower AUC than RankSVM_{CNN} on the test set, which verifies that the motivation of the top-rank learning method is to maximize the Pos@Top, not AUC
- 3) TopRank CNN obtained a higher Pos@Top than the TopPush_{CNN} that used a fixed feature representation given by CNN_{soft}. This revealed the superiority of TopRank CNN on the maximization of Pos@Top by end-to-end learning

- 4) TopRank CNN also achieved the best AUC, although its main objective was not to maximize AUC but Pos@Top. In fact, TopRank CNN achieves higher AUC values than RankSVM_{CNN} (which just tries to maximize AUC). This suggests that the feature representation in the training process that boosts Pos@Top may give a positive effect to maximize AUC on the test phrase
- 5) Comparing TopRank CNN to CNN_{pos@top}, p -norm relaxation by Eq. (4.1) is effective in reducing the overfitting to the training samples. In fact, TopRank CNN achieves a similar Pos@Top (0.0960) on the test samples when p is increased to 64
- 6) CNN_{focal} achieves a better performance than CNN_{soft} and thus shows its strength in the unbalanced task; however, its AUC, as well as Pos@Top does not exceed TopRank CNN

To observe the ability of TopRank CNN to maximize Pos@Top, the t-distributed stochastic neighbor embedding (t-SNE) [83] was applied to visualize the deep features and present the related representation space. The t-SNE results of each method (with the optimal hyper-parameter) on the CIFAR-10 dataset are shown in Fig. 4-3. The red dots depict the absolute top samples given by each method, whereas the green dots depict positive samples that are not selected as the absolute top.

Among all the methods, TopRank CNN has more absolute top samples around the edge of the distribution owing to the allocation of larger rank scores to positive samples. The other methods also yield a reasonable separation between positive and negative samples. However, most of them (except for CNN_{pos@top}), cannot have more absolute top samples.

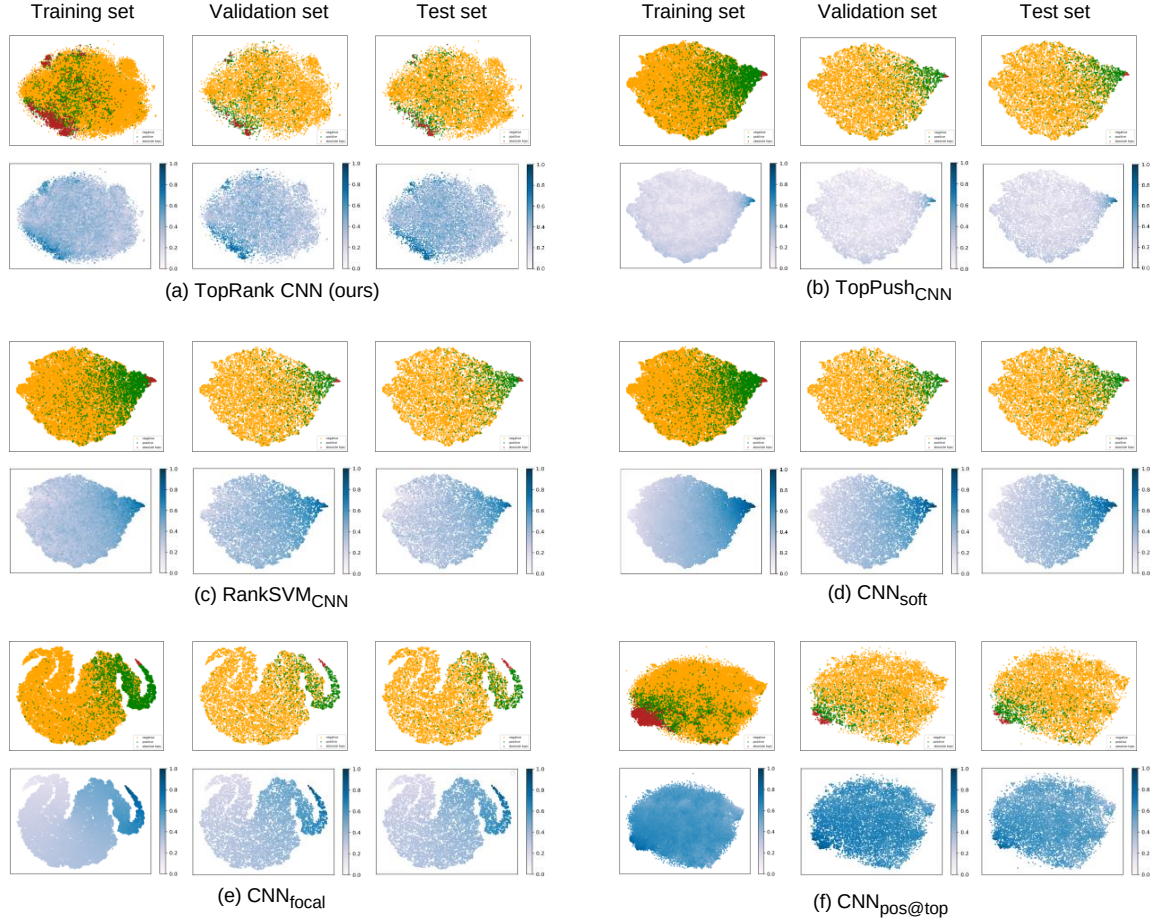


Figure 4-3: t-SNE visualizations of the training, validation, and test sets of the unbalanced CIFAR-10 dataset. For each method, the upper row shows positive (green and red; red denotes *the absolute top*) and negative (orange) sample distribution and the lower shows their ranking scores (darker color is higher). For the classification methods, such as CNN_{soft}, their soft-max values are used as the ranking scores. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5 Experiment on a Medical Image Diagnosis Task

As noted in Section 1.2.2, more absolute top samples are crucial for medical image diagnosis for more reliable decisions. In this section, the proposed method is applied to a mammogram image diagnosis task to observe the practical performance of the proposed method. The experimental results show that the proposed method TopRank

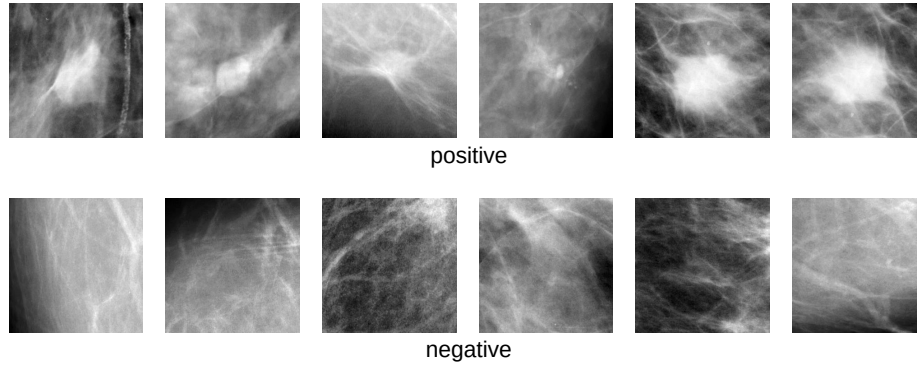


Figure 4-4: Examples from the DDSM Mammography dataset. The upper row shows the masses or calcification images (positive) and the bottom row shows the normal images (negative).

CNN can have a higher Pos@Top and is also very robust to this unbalanced diagnosis task.

4.5.1 Dataset

The DDSM mammography open-source dataset ¹ was used to evaluate the performance of the proposed method in practical medical-image-based diagnosis tasks. Fig. 4-4 shows several samples from this dataset. The DDSM mammography dataset is a large-scale unbalanced binary mammogram image dataset that consists of negative samples (normal images) from the DDSM dataset [84] and positive samples (abnormal images) from the CBIS-DDSM dataset [85]. The sample images are pre-processed and converted to 299×299 pixels by extracting the regions-of-Interest (ROIs). It contains 55,885 training samples, of which 14% are positive samples, and the remaining 86% are negative samples. Accordingly, the ratio of the positive and negative samples on the training set is 1 : 7. The test set consists of 15,364 images, which are constructed by 2,004 masses or calcification images (positive) and 13,360 normal images (negative).

¹<https://www.kaggle.com/skooch/ddsm-mammography>

Table 4.2: $Pos@Top$ and AUC of TopRank CNN (ours) and comparative methods on the DDSM Mammography dataset.

(a) Result on the original DDSM Mammography test set (the ratio of positive and negative samples is 1 : 7).

Method	Feature	Algorithm	Loss	Hyper-param.	AUC	Pos@Top
TopPush _{CNN}	CNN _{soft}	TopPush	Eq.(2.11)	$\lambda = 0.1$	0.9317	0.1362
RankSVM _{CNN}	CNN _{soft}	RankSVM	Eq.(2.2)	$C = 10$	0.9799	0.1626
CNN _{soft}	end-to-end	CNN	Softmax	Epoch=100	0.9780	0.1043
CNN _{focal}	end-to-end	CNN	Focal	Epoch=100	0.9713	0.1290
TopRank CNN (ours)	end-to-end	CNN	Eq.(4.1)	$p = 32$, Epoch=30	0.9791	0.3183

(b) Result when the extremely unbalanced DDSM Mammography training samples (whose the ratio of positive and negative samples is 1 : 97) are used.

Method	Feature	Algorithm	Loss	Hyper-param.	AUC	Pos@Top
TopPush _{CNN}	CNN _{soft}	TopPush	Eq.(2.11)	$\lambda = 0.1$	0.8544	0.0115
RankSVM _{CNN}	CNN _{soft}	RankSVM	Eq.(2.2)	$C = 0.01$	0.9014	0.0050
CNN _{soft}	end-to-end	CNN	Softmax	Epoch=100	0.9319	0.0000
CNN _{focal}	end-to-end	CNN	Focal	Epoch=100	0.8315	0.0000
TopRank CNN (ours)	end-to-end	CNN	Eq.(4.1)	$p = 8$, Epoch=50	0.9523	0.1023

4.5.2 Comparative Methods

TopRank CNN is evaluated against four comparative methods, including CNN_{soft}, RankSVM_{CNN}, TopPush_{CNN}, and CNN_{focal}. In this experiment, the backbone architecture of CNN is the fully convolutional network (FCN) applied to recognize abnormalities from the DDSM mammography dataset ² that comprised 10 convolutional layers. As proved in the case of the unbalanced CIFAR-10 dataset, the loss function of Eq. (2.11) with larger p approaches to the TopRank loss (Eq. 4.1) may lead to overfitting problems. Thus, the comparative experiment is not conducted with CNN_{pos@top} in this part.

Five-fold cross-validation was used for model validation based on the use of the same settings of hyper-parameters as reported in the reference literature [18, 3]. According to the validation, λ was selected from $\{10^{-2}, 10^{-1}, 1, 10^1, 10^2\}$ and C was selected from $\{10^{-2}, 10^{-1}, 1, 10^1, 10^2\}$. In consideration of the numerical overflow problem about the p -norm with larger p values, in these experiments, p was selected from the set of $\{2, 4, 8, 16, 32\}$.

²<https://www.kaggle.com/skooch/fcn-for-detecting-abnormalities-in-mammograms>

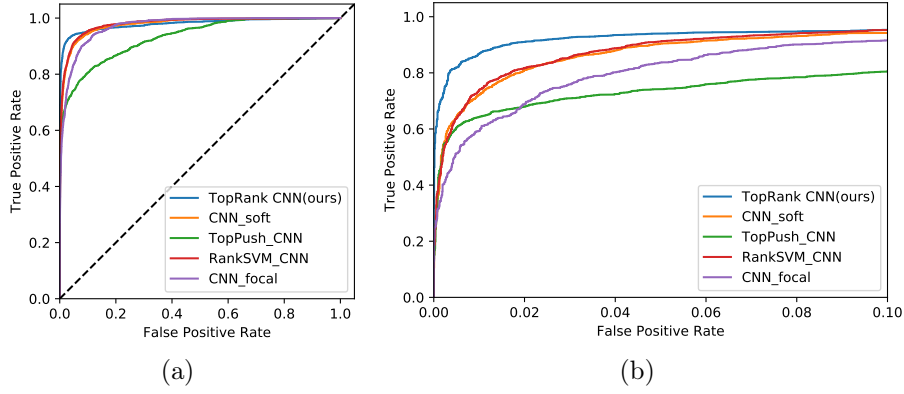


Figure 4-5: (a) ROC of TopRank CNN (ours) and comparative methods on the test set of the DDSM mammography dataset. (b) The leftmost vertical part of (a).

4.5.3 Results on the Medical Image Diagnosis Task

The quantitative evaluation results are shown in Table 4.2(a). For Pos@Top, TopRank CNN largely outperforms CNN_{soft} , $\text{TopPush}_{\text{CNN}}$, $\text{RankSVM}_{\text{CNN}}$, and $\text{CNN}_{\text{focal}}$, as expected. Additionally, the AUC of the TopRank CNN was almost the same as $\text{RankSVM}_{\text{CNN}}$ (which was designed to maximize AUC). Fig. 4-5 shows the ROC of each method. As explained (Fig. 1-2), the leftmost vertical part of the ROC was Pos@Top. This is very important, especially for medical diagnosis, as Pos@Top presents the ratio of absolute top samples (which are very reliable positive samples). These ROC plots also support the fact that the proposed method can find more absolute top samples than the other methods.

Fig 4-6 shows the top-10-ranked medical image samples among the test set. Note that all of them are positive samples (according to their ground-truth images) and belong to the absolute top.

Fig. 4-7 shows the t-SNE visualization of the sample distributions in the feature spaces of individual methods. Just like the unbalanced CIFAR-10 dataset, it is confirmed that TopRank CNN could form a larger area for absolute top samples than the other methods.

Finally, the effects of p -norm relaxation on the maximization of Pos@Top was

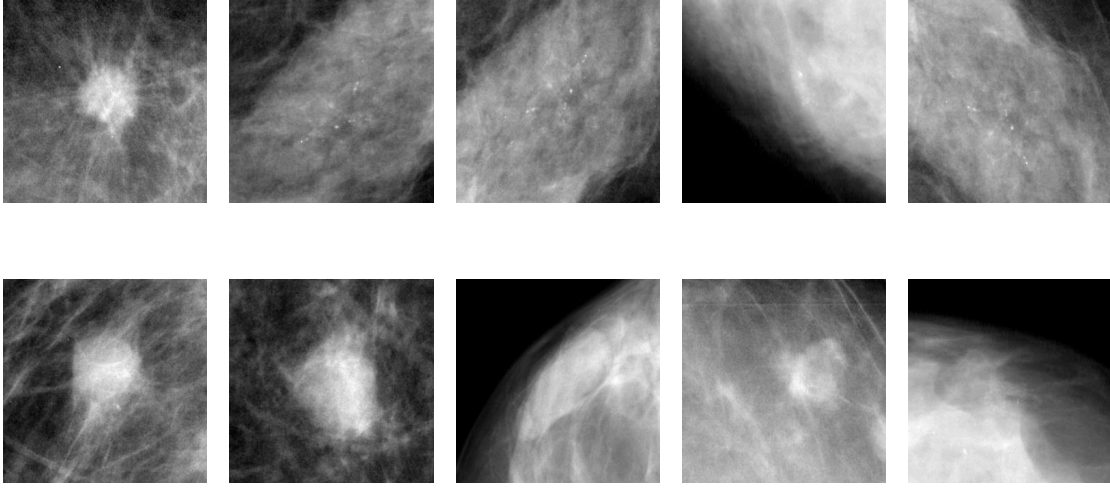


Figure 4-6: Top-10 ranked samples in the test set by the TopRank CNN (ours). All of these samples are abnormal (positive) images according to their ground-truth.

observed. Fig. 4-8 shows Pos@Top for the validation set with different p values. Given that a larger p will mimic the original loss function (Eq. 4.1) to maximize Pos@Top, Pos@Top becomes larger along with p . However, Pos@Top is almost saturated after $p = 16$. This fact is very favorable for the proposed TopRank CNN because its p -norm relation loss (Eq. 2.11) suffers from numerical overflows at larger p values.

4.5.4 Results on Extremely Unbalanced Data

The experiments described have proved the superiority of TopRank CNN for solving medical image diagnosis tasks. Specifically, the minorities (the positive samples) are not ignored but rather emphasized based on its training process. Finally, larger absolute top samples could be obtained. A new question arises: what will happen if the minorities are more minor? In other words, is it worthy to observe the performance of TopRank CNN on extremely unbalanced data?

For this observation, an extremely unbalanced dataset was created from the original DDSM mammography dataset. Specifically, randomly picked 500 positive images and all the negative images from the training set are used to build the new training

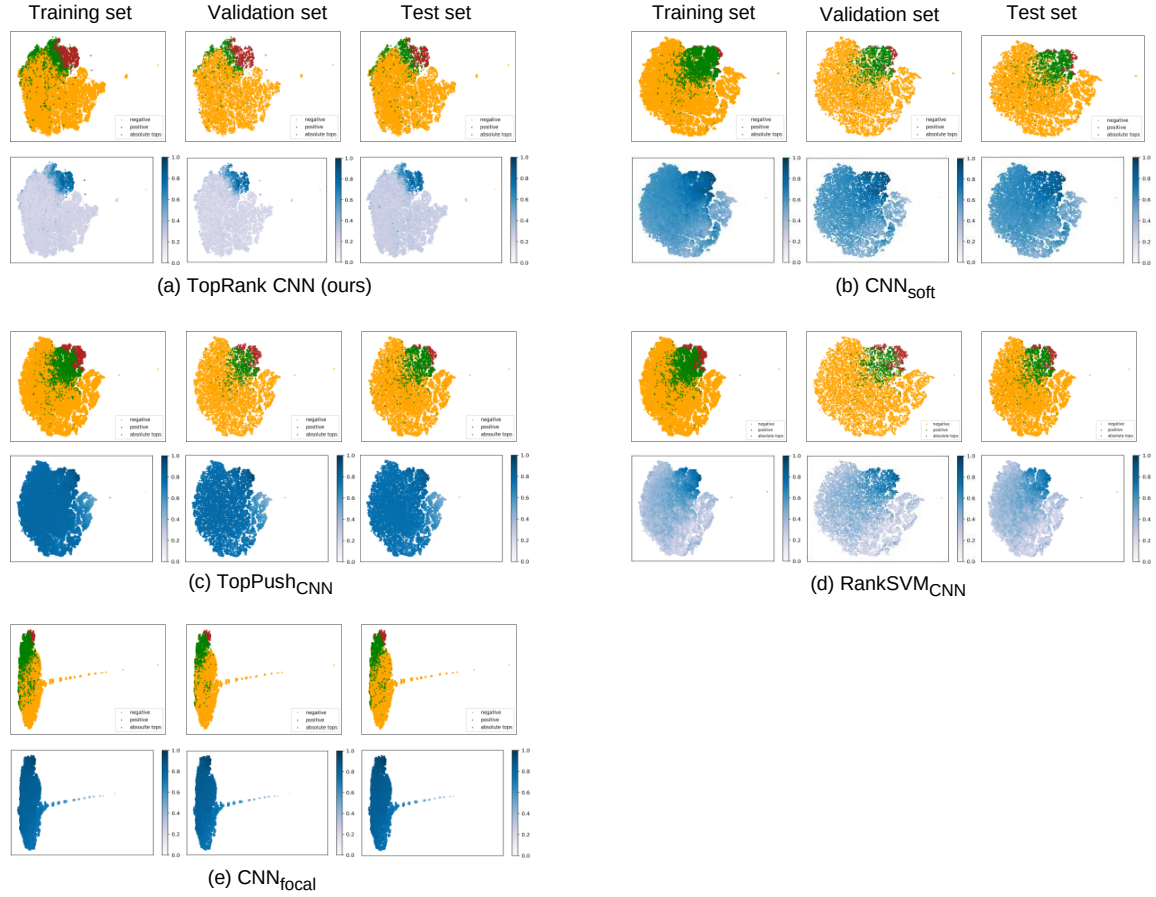


Figure 4-7: t-SNE visualization of the training, validation, and test sets of the DDSM Mammography dataset. For each method, the upper row shows positive (green and red; red denotes *the absolute top*) and negative (orange) sample distribution and the lower shows their ranking scores (darker color is higher). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

set. This dataset was composed of 1.02% positive samples and 98.98% negative samples. The ratio of positive and negative samples of training data was changed from 1 : 7 to 1 : 97. The original test set was then used to test the performance of each method. Except for the training data and class weight hyper-parameter, the same experimental settings with the previous experiment was applied, including comparative methods and evaluation metrics, as well as the CNN architecture.

Table 4.2(b) shows the test results of Pos@Top and AUC, and it confirms that the

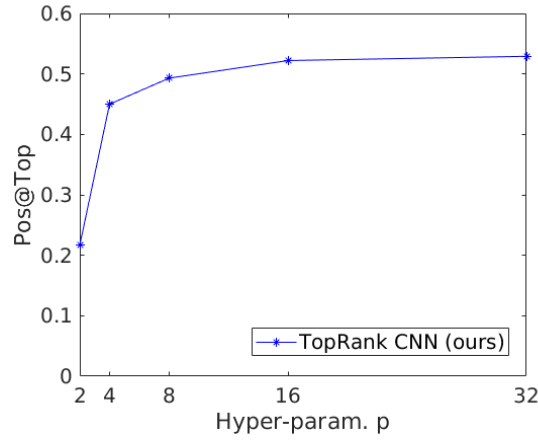


Figure 4-8: Pos@Top of TopRank CNN (ours) with different p on the validation set of the DDSM Mammography dataset.

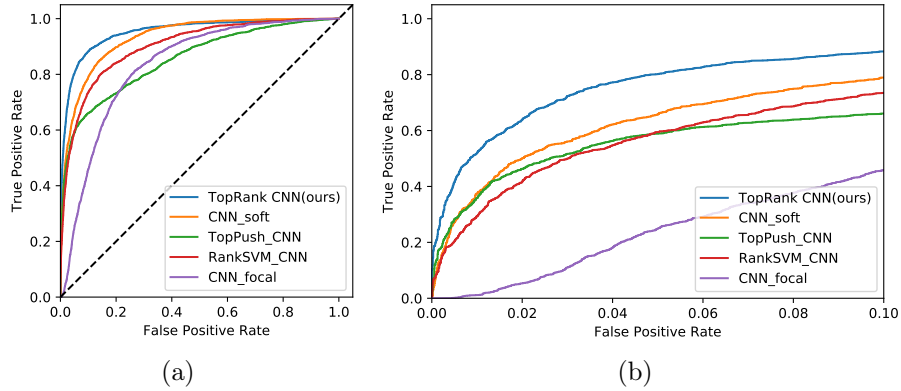


Figure 4-9: (a) ROC of TopRank CNN (ours) and comparative methods on the test set when the extremely unbalanced DDSM mammography data is used for training. (b) The leftmost part of (a).

TopRank CNN still outperforms the other methods based on its Pos@Top and AUC values. Although the Pos@Top of TopRank CNN becomes smaller than the original dataset owing to this extremely unbalanced condition, its Pos@Top is far larger than the other methods.

This superiority is also confirmed by the ROC shown in Fig. 4-9. The lower ROC outcomes of the comparative methods indicates their difficulties to identify the reliable, positive samples (i.e., absolute top samples).

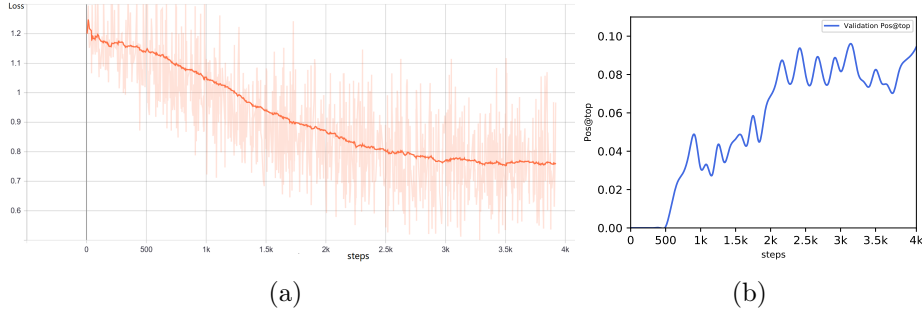


Figure 4-10: (a) The training loss curve of the proposed TopRank CNN with $p=8$ when the extremely unbalanced DDSM Mammography training samples (whose the ratio of positive and negative samples is 1 : 97) are used. (b) The corresponding Pos@Top curve on validation data at different steps.

For $\text{CNN}_{\text{focal}}$, different hyperparameters were tried for experiments, but significant improvements were not obtained. The corresponding results shown in the paper were achieved by setting the gamma and alpha to 2 and 0.25 due to the best performance of Focal loss in the literature [82]. (Note that although a further hyperparameter search was conducted, the best hyperparameter values were found as $\alpha=0.99$ and $\gamma=0$; this means that the focal loss is reduced to the softmax loss for this dataset. AUC and Pos@Top for the case were 0.8823 and 0.002, respectively.)

To ensure the convergence of TopRank loss in such an extreme unbalanced situation, the training loss curve and validation performance of the proposed TopRank CNN were presented in Fig. 4-10. When steps reached 4k, the training loss value became stable and the Pos@Top on validation data also reached the peak and stabilized in a small fluctuation range.

4.6 Summary

This chapter proposed *Top-rank convolutional neural network* (TopRank CNN) that maximized the number of absolute top samples that were very reliable positive samples. This is the first discussion of a top-rank learning approach with representation learning in an end-to-end way. Featuring a combination of top-rank learning and

representation learning with p -norm relaxation, TopRank CNN maximized the number of absolute top samples while it formed the appropriate feature space. As shown by the experimental results on medical image-based diagnosis tasks, TopRank CNN outperformed other top-rank learning methods and conventional deep CNNs based on the higher ratios of absolute top samples (i.e., Pos@Top) and higher AUC values. The robustness to unbalanced tasks was also proved experimentally. One point needs to be observed in future work: there are still positive samples not assigned into Pos@Top in TopRank CNN. Top-ranked negative samples determine these positive samples so that if I can find a more global top-ranked negative may improve the performance of Pos@Top. That means mini-batch may affect the performance of TopRank CNN on maximizing Pos@Top. I will observe the impact of the mini-batch in the future.

Chapter 5

Robust Top-rank learning for Medical Image Analysis

5.1 Purpose

In this chapter, a robust top-rank learning algorithm is proposed to cope with the negative effect of outliers for the conventional top-rank learning, with an unsupervised outlier estimation technique called local outlier factor (LOF). Introduction of LOF can weaken the effect of the negative outliers and thus increase the stability of the learned ranking function. Moreover, robust top-rank learning with representation learning is combined by a deep neural network (DNN). In the later experiments, I will use two artificial datasets and a medical image dataset to verify the robustness of the proposed method to outliers.

First, it gives the review for related work about outlier estimation methods. Then, it discusses the formulation of Robust Top-rank learning and its combination with representation learning. Finally, it summarizes the experimental results and gives a conclusion for robust top-rank understanding.

5.2 Related Work about Outlier Estimation

In this section, the related work about outlier estimations is briefly introduced. To evaluate the degree of outliers, many unsupervised outlier estimation methods (, or unsupervised anomaly estimation methods,) are extensively studied [86, 87, 88]. In general, these studies assume that inliers are located in dense areas while outliers are not. Therefore, the outlier estimation can be characterized by their density estimation method, such as statistical methods [86, 87, 88], neighbor-based methods [33, 89], and reconstruction-based methods [90, 91]. Robust model fitting algorithms are also prevalent in many computer vision applications where the dataset is contaminated with outliers [92] which aim to learn robust models with consensus maximization by maximizing the number of inliers and achieve efficient performance. Different from robust fitting methods, the purpose of this work is to weaken the effect of outliers in top-rank learning.

In this study, LOF [33] is employed, one of the most common unsupervised outlier estimation methods, for the proposed method. The strong characteristic of LOF is that it can evaluate the local anomaly [87], which is often overlooked by other common methods, such as the k NN-based outlier estimator. Note again that the proposed method can employ other outlier estimators by virtue of its flexibility.

Recently, outlier estimation methods have been employed in DNN learning tasks. Specifically, these methods are used for sample selection or sample reweighting for classification tasks [93, 94, 95]. It should be emphasized that top-rank learning is further sensitive to outliers than classification tasks, as noted in Section 1.2.3 and this study is the first attempt to introduce an outlier estimation method to the top-rank learning task.

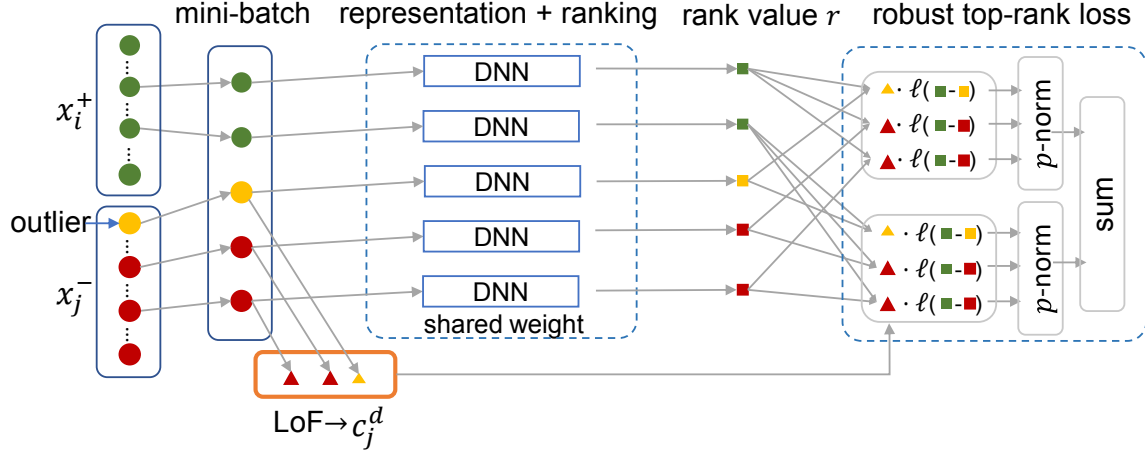


Figure 5-1: The overview of the proposed method, robust top-rank learning.

5.3 Robust Top-Rank Deep Neural Network

5.3.1 Top-Rank Learning Robust to Outlier

Top-rank learning aims to learn a real-valued ranking function that gives the positive samples higher rank scores than top-ranked negative samples. However, top-rank learning is drastically degraded just by a single negative outlier in training data as the Fig. 1-3 (b) is shown in Section 1.2.3. The degradation appears as not only the decrease of absolute tops (i.e., “doubtlessly positive” samples) but also the instability of the ranking function. The latter degradation is more serious, especially for using the learned ranking function for test samples.

Therefore, this thesis newly introduces the LOF into the objective function for weakening the effect of outliers on the ranking function by top-rank learning in Eq. (2.11). Specifically,

$$\mathcal{J}_{\text{RobustTopRank}} = \frac{1}{m} \sum_{i=1}^m \max_{1 \leq j \leq n} c_j^d \cdot \ell(r(\mathbf{x}_j^-) - r(\mathbf{x}_i^+)), \quad (5.1)$$

The negative sample $\mathbf{x}_j^-$ that gives the maximum in Eq. (5.1) is the top-ranked negative. The $r(\cdot)$ denotes the ranking function. The function ℓ is chosen as $\ell(z) =$

$\log(1 + e^{-z})$ in later experiments, and c_j^d is the weight given by

$$c_j^d = \left(\frac{1}{\max(\text{LOF}_k(\mathbf{x}_j^-), 1)} \right)^d. \quad (5.2)$$

Here, $\text{LOF}_k(\mathbf{x}_j^-)$ gives LOF value of $\mathbf{x}_j^-$. (The role of the hyperparameters $d \in \mathbb{N}^+$ and $k \in \mathbb{N}^+$ will be explained later.) A higher degree of anomaly gives a larger LOF value and thus a smaller c_j^d . Thus, the effect of outlier samples will be automatically weakened during the training of r .

There are two points to note. First, in Eq. (5.1), the weight c_j^d is imposed only on the negative samples. As discussed in Fig. 1-3, the purpose is to avoid that a negative outlier becomes the top-ranked negative. Therefore, this study introduce the weight so that the effect of negative outliers is weakened. Second, $\text{LOF}_k(\mathbf{x}_j^-)$ (or, equivalently, the weight c_j^d) is pre-calculated in the original feature space of the training samples Ω and fixed during the training process. In other words, although introduce representation learning will be introduced as discussed in Section 5.3.2, LOF is not evaluated in the learned feature space. This is because the proposed method needs to avoid the unexpected representation learning result where all negative samples become outliers for minimizing the objective function with smaller c_j^d .

The hyperparameter d controls the degree of outlier weakening. A larger d will weaken outliers more strongly. Note that $c_j^d \in (0, 1]$ for any d . The other hyperparameter k defines the number of nearest neighbors to calculate LOF. In general, the hyperparameter k should be determined by watching the data distribution. In the later experiment with medical image data, $k = 30$ is set by observing the training samples.

5.3.2 Combination with Representation Learning

This study combines the above robust top-rank loss (Eq. 5.1) with representation learning, which is realized by a DNN, such as CNN. Eq. (5.1) contains the max

operation. It prevents us from the end-to-end collaborative training of the ranking function r and the feature representation. Therefore, the p -norm relaxation technique is introduced to replace the max operation with a differentiable operation:

$$\mathcal{J}_{\text{Relaxed}} = \frac{1}{m} \sum_{i=1}^m \left(\sum_{j=1}^n (c_j^d \cdot \ell(r(\tilde{\mathbf{x}}_i^+) - r(\tilde{\mathbf{x}}_j^-)))^p \right)^{\frac{1}{p}}, \quad (5.3)$$

where $\tilde{\mathbf{x}}$ denote the feature vector given by DNN. Although $p \rightarrow \infty$ will give a better approximation of Eq. (5.1) theoretically, it will cause the overflow problem in practice. Therefore, $p = 32$ is set at maximum in the later experiment.

As shown in Fig. 5-1, the DNN can be trained, which gives the ranking function r and the representation $\tilde{\mathbf{x}}$ in an end-to-end manner. DNN in this figure plays both roles; its earlier layers extract appropriate features $\tilde{\mathbf{x}}$, and the final layers determine the rank value $r(\tilde{\mathbf{x}})$. This end-to-end training utilizes its representation ability for obtaining more absolute tops while increasing the stability of the ranking function by weakening the effect of negative outliers.

In practice, the loss of Eq. (5.3) is calculated for each mini-batch, and the mini-batch should consist of more negative samples than positive samples. It is because the purpose of top-rank learning is to give higher rank values to positive samples than the higher-ranked negative samples; therefore, it is better to increase the probability that the mini-batch contains higher-ranked negatives by increasing the ratio of negative samples. Note that, different from the classification task, class imbalance does not matter for the ranking task. This is because positive and negative samples are always paired in the objective function, and their effect is equivalent.

5.4 Experiment on Artificial Datasets

To show the robustness of the proposed method to outliers, experiments are conducted with two artificial datasets; one dataset contains 100 one-dimensional train samples,

and the other has 100 two-dimensional train samples. The data distributions of their training sets are shown in Figs. 5-2 (a) and 5-3 (a), respectively.

In this experiment, the convolutional layers are not used for representation learning; directly use the original sample in Eq. (5.3) and use four fully connected layers to model the ranking function r . Hyper-parameters $d = 100$ and $k = 15$ are set (using the same-sized validation set, which is also used for determining the number of the training epochs). Different $p \in \{2, 4, 8, 16, 32\}$ are examined. For comparison, the original top-rank learning is realized by setting $c_j^d = 1, \forall j$.

Fig. 5-2 (b) plots the learned ranking functions by the original top-rank learning and the proposed method, respectively. It proves that the ranking functions by the original top-rank learning show a big drop around the negative outlier (at $x^- = 1$). In contrast, the proposed method does not show such a drop successfully by weakening the effect of the negative outlier¹. I also try to observe the performance of the learned ranking function when the model suffers from outliers of the positive class by this experiment. In Fig. 5-2 (a), the leftmost (green) point denotes the positive outlier. For the original Top-rank learning method, the outlier of positive sample has a lower rank score than these positive samples in the middle position, which may produce a severe effect for the ranking function or final results of Pos@Top, as Fig. 5-2 (b) is shown. In contrast, the robust top-rank learning method assigns a similar ranking score to the positive outlier with these positive samples in the middle position. It means that the robust top-rank learning may weaken the effect of positive outliers in the training process.

Fig. 5-3 (b) shows the results of the two-dimensional artificial data. The rank function is represented as a heatmap between dark-blue (lower rank) and yellow (higher rank). Like the one-dimensional case, the original top-rank learning suffers from a significant drop around the negative outlier; in other words, even a single outlier can

¹The ranking function r only specifies the relative relationship between two samples (such that $r(\mathbf{x}^+) > r(\mathbf{x}^-)$). Thus, its absolute value is not important. Therefore, the curve shape is more important when Fig. 5-2 (b) is observed.

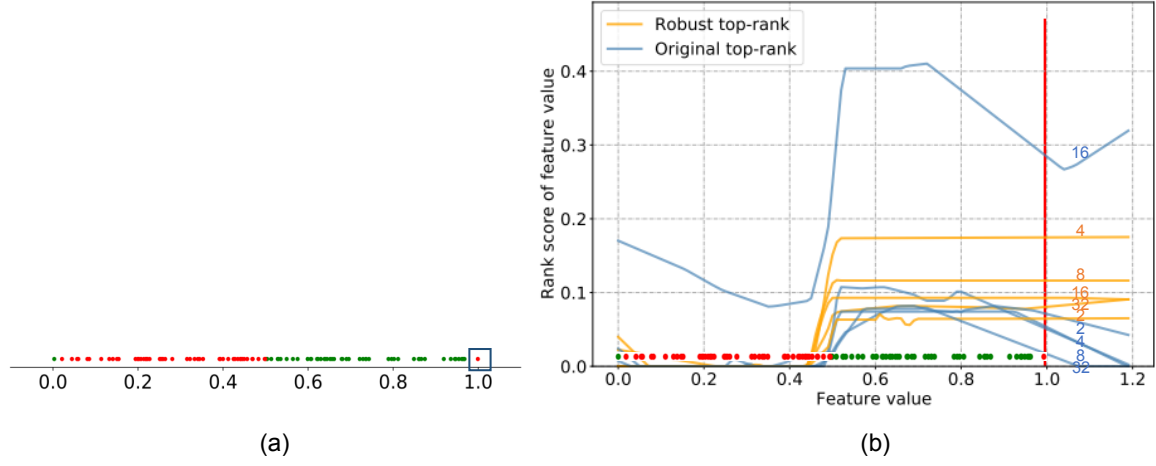


Figure 5-2: (a) One-dimensional artificial dataset. The blue box indicates the outlier. (b) Results of the original top-rank learning and the robust top-rank learning. The red vertical red bar in (b) indicates the negative outlier location. A digit on a curve indicates the value of p .

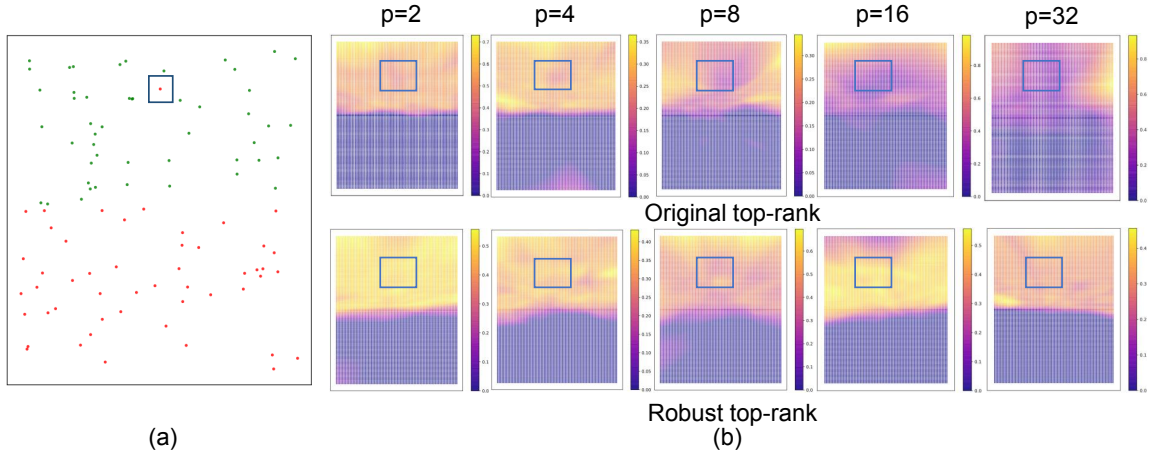


Figure 5-3: (a) Two-dimensional artificial dataset. (b) Results of the original top-rank learning and the robust top-rank learning. In the heatmap representation, purple indicates a lower ranking score, and yellow indicates a higher ranking score. The blue box indicates the outlier location.

degrade the ranking function seriously. However, the outlier effect is further weakened by the proposed method.

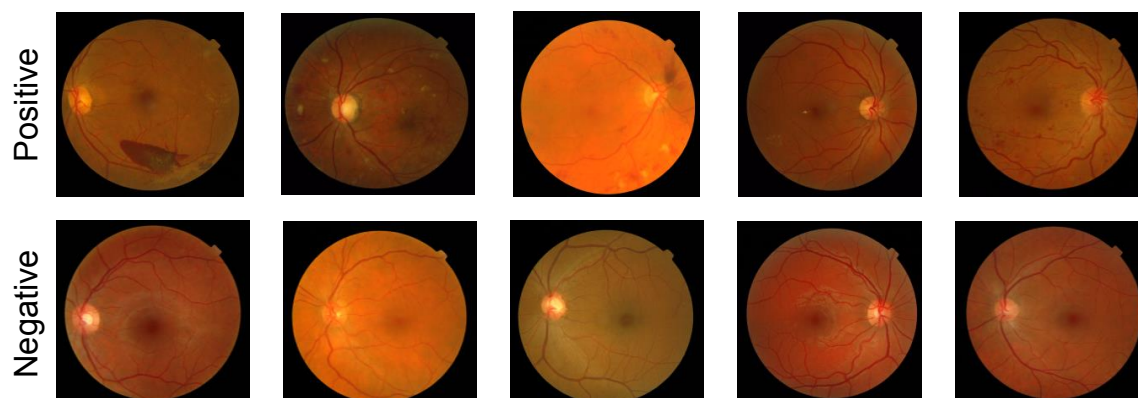


Figure 5-4: Examples from the Messidor dataset. The first and the second row respectively show the disease images (positive) and the normal images (negative).

5.5 Experiment on Medical Image Diagnosis

5.5.1 Dataset

As a practical application of the proposed method, binary diabetes retinopathy (DR) ranking experiment is conducted on the Messidor dataset [96], which is public on website ². Fig. 5-4 shows several samples from the Messidor dataset. As noted in Section 1.2.3, the datasets of medical image analysis tasks often contain outliers due to measurement error or labeling error, and the original top-rank learning will become unstable by the outliers.

Recently, He et al. [1] proposed a binary DR grade classification method, called CABNET, and achieved state-of-the-art performance. Therefore, all image samples are converted into the CABNET features in advance. In this experiment, the same data division was followed as [1]. Consequently, 501 positive samples and 699 negative samples are used in our experiments.

Fig. 5-5 shows the t-SNE visualization of the CABNET features of the train and test samples. As expected, several negative outliers are found in the positive sample distribution, probably because of the labeling difficulty.

²<https://www.adcis.net/en/third-party/messidor/>

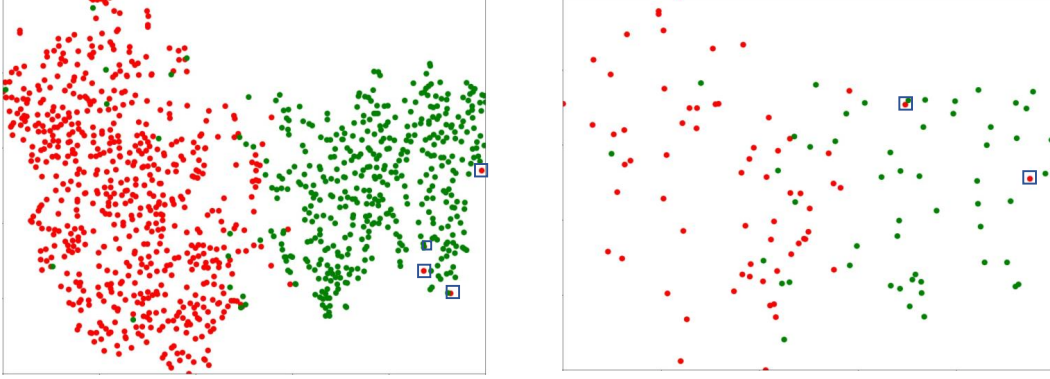


Figure 5-5: t-SNE visualizations of initial feature vectors (by CABNET [1]) of the training (left) and test (right) samples of the Messidor dataset. The red dots denote negative samples, and the green dots denote positive samples. The red dots with the blue box indicates the negative outliers.

5.5.2 Model Setup

A CNN is used as the DNN of Fig. 5-1. Each CNN has three convolutional layers and two fully connected layers. ReLU and max-pooling operation and batch-normalization followed each convolutional layer. Since the CABNET features are used as the input features, such a shallow CNN is sufficient. In training, the stochastic gradient descent (SGD) optimization algorithm was utilized. As noted in Section 5.3.2, a very unbalanced mini-batch with 5 positive samples and 45 negative samples is used for training.

5.5.3 Comparative Methods

To measure the performance of the proposed method, comparative experiments are conducted with CABNET, RankNet[38, 97], and TopRank CNN[98]. RankNet is a very famous learning to rank algorithm that combines representation learning. For a fair comparison, RankNet and TopRank CNN use the same architecture and the same input (CABNET feature vectors) as the proposed method. In other words, they are different only at their loss function.

Table 5.1: Pos@Top on the Messidor dataset. For a fair comparison, the same network architecture have been used for all the methods.

Method	Pos@Top $\uparrow$	
	Train	Test
CABNET [1]	0.4286	0.5379
RankNet[38]	0.0803	0.2449
TopRank CNN[98]	0.3368	0.5896
Ours	0.5433	0.6524

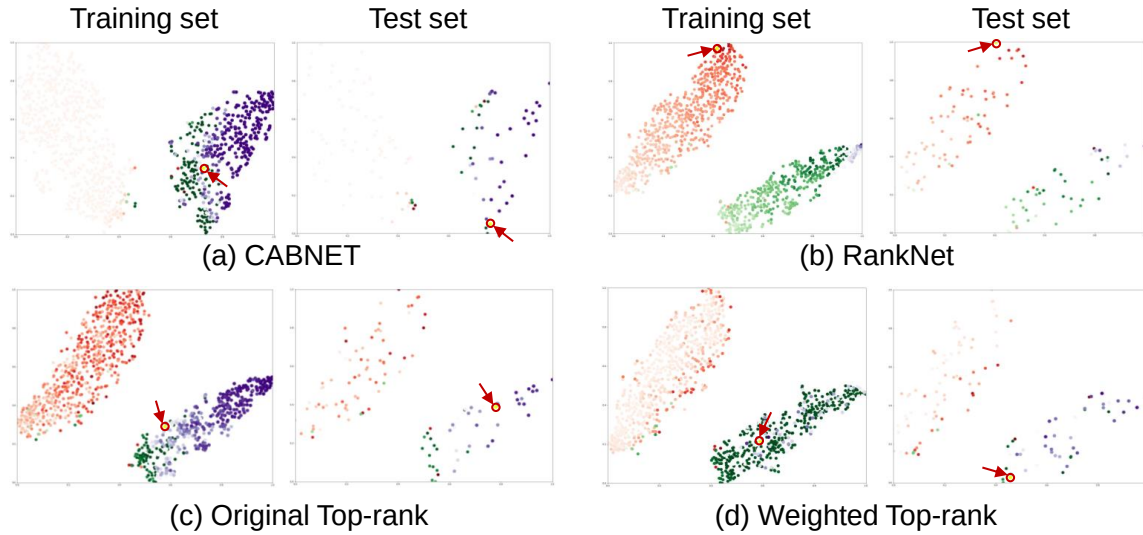


Figure 5-6: t-SNE visualizations results of the comparative methods on the Messidor dataset. Red, green, and purple dots denote negative, positive, absolute top (i.e., doubtlessly positive) samples, respectively; yellow dots with a red circle (pointed by a red arrow) denote the top-ranked negative sample.

The ten-folds cross-validation method is used to tune the hyperparameters. The partitions of each fold data are followed the CABNET [1]. For the proposed method, hyperparameters p and d are selected from $\{2, 4, 8, 16, 32\}$ and $\{1, 10, 100\}$, respectively. TopRank CNN also employs p -norm relaxation for its top-rank loss and thus has the same hyperparameter p ; its p is also optimized in the same way.

5.5.4 Quantitative Evaluation by Pos@Top

Table 5.1 presents the Pos@Top of the comparison method, which are average among ten-fold test data. (The soft-max values are used to calculate the Pos@Top for the CABNET.) It should be emphasized that the proposed method has *no guarantee* of achieving higher Pos@Top for *the training set*; Fig. 1-3 (c) depicts the extreme case where the weakened negative outlier is ranked at the top, and thus there is no absolute top (Pos@Top=0). However, it should be more emphasized that the ranking function of Fig. 1-3 (c) is more suitable than Fig. 1-3 (b) because it is similar to Fig. 1-3 (a), which shows the ranking function in the absence of the negative outliers. Therefore, it is expected that the proposed method will give a higher Pos@Top for *the test set*³.

As shown in Table 5.1, the proposed method successfully outperforms the comparative methods for *the test set*. Especially, the comparison with TopRank CNN proves that the proposed top-rank learning method could weaken the effect of the outliers and get more stable results. Since the proposed method shows higher Pos@Top, it can derive more reliable positive (i.e., doubtlessly positive) samples — this is very beneficial for medical diagnosis tasks.

The t-SNE visualization of a one-fold dataset distribution on the learned feature space is shown in Fig. 5-6. The proposed method achieved better performance on the test set than the comparison method can be confirmed. As noted above, the proposed method cannot have many absolute tops in *the training set* due to the outliers. However, the learned ranking function is still reasonable and thus could have more absolute tops in *the test set*.

5.5.5 Quantitative Evaluation of Ranking stability

Another quantitative evaluation of the stability of top-rank learning methods is also conducted by using the Kendall tau distance and Spearman correlation coefficient.

³If negative outliers can be removed in the evaluation, the ranking function of Fig. 1-3 (c) will achieve more Pos@Top than Fig. 1-3 (b), even for the training set.

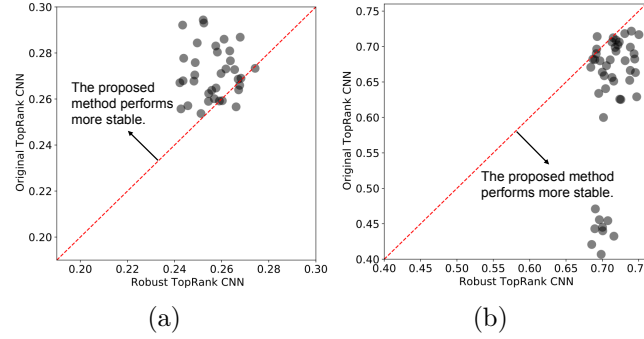


Figure 5-7: (a) Kendall Tau distance and (b) Spearman correlation coefficient. Those values by TopRank CNN and the proposed method are depicted in vertical axes and horizontal axes, respectively.

Among the ten-folds (for the cross-validation), some folds contain a certain negative outlier, and the others do not. If the ranking method is sensitive to the outlier, the ranking of the samples might change drastically. If not, the ranking will be almost the same. The Kendall tau distance and Spearman correlation coefficient can evaluate the dissimilarity and similarity of two ranking results of the same dataset, respectively.

Fig. 5-7 shows (a) the Kendall tau distance and (b) Spearman correlation coefficient between every two folds for the proposed method and the original TopRank CNN on the ranking. Since it has 45 ($=_{10}C_2$) fold-pairs, both of (a) and (b) have 45 dots. In (a), the proposed method almost always shows a smaller distance. In (b), the proposed method almost always shows a larger correlation coefficient. Therefore, it can be concluded that the proposed method is more stable by weakening the effect of the outliers.

5.6 Summary

This chapter proposed a novel top-rank learning method that can weaken the effect of outliers in the training process by using an unsupervised outlier detection technique. To my best knowledge, it is the first proposal to realize robust top-rank learning. Moreover, we combine the robust top-rank learning framework with representation

learning in an end-to-end way. Experiments on artificial datasets and medical image datasets demonstrate the robustness of the proposed method for outliers. Benefits from the robustness to outliers, the proposed method obtains more reliable positive samples (higher Pos@Top) in the test set than the comparison methods on medical image diagnosis tasks. Also, the stability of the learned ranking function is confirmed experimentally. In the future, I will pay more attention to positive outliers and observe their effect. Further analysis of the characteristic of robust top-rank learning will be given, and the applications on other datasets will be observed.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, I introduce the merits and shortcomings of learning to rank, especially for top-rank learning, which aimed to learn a real-valued ranking function to provide higher ranking scores to positive samples than negative samples. This property enables learning to rank robust to data imbalance problems. More importantly, top-rank learning can maximize the number of absolute top samples that were very reliable positive samples. The main purpose of this thesis is to discuss the combination of top-rank learning with representation learning (neural networks) in an end-to-end way. In addition, it is demonstrated the negative effect of outliers on top-rank learning and proposed robust top-rank learning to weaken the effect of outliers in the training process.

As preliminary research of top-rank learning, a conventional learning to rank approach-RankSVM is first introduced, which has several good properties for solving the data imbalance problem in signature verification tasks. Especially, it theoretically guarantees the robustness to the imbalanced task without any parameters compared to standard SVM. Extensive experiments were conducted on signature verification and retrieval tasks. The results demonstrated that: (1) RankSVM could give the

most reliable genuine samples as the top-ranked samples among all genuine samples, but standard SVM often fails to get them. (2) RankSVM could achieve significantly higher performance than the standard SVM for both signature retrieval and signature verification tasks, especially in a tough situation where the positive and the negative classes are heavily overlapping. It also revealed that RankSVM could be affected by the distributions of the positive and the negative samples.

After that, top-rank learning was employed for medical image analysis tasks. Top-rank learning aimed to maximize the number of *absolute tops* that are very reliable positive samples, which enabled top-rank learning very cater to the medical image analysis task that required a reliable decision. In this task, I propose the *Top-rank convolutional neural network* (TopRank CNN) to combine top-rank learning and representation learning. It was the first discussion of a top-rank learning approach with representation learning in an end-to-end way by p -norm relaxation. TopRank CNN maximizes the number of absolute top samples and formed the appropriate feature space. The experimental results demonstrated that TopRank CNN outperformed other top-rank learning methods and conventional deep CNNs based on the higher ratios of absolute top samples (i.e., Pos@Top) and higher AUC values on medical image-based diagnosis tasks. The robustness to unbalanced tasks was also proved experimentally.

To solve the problem of top-rank learning of the sensitivity to outliers, a robust top-rank learning method is proposed to weaken the effect of outliers in the training process by using an unsupervised outlier detection technique-local outlier factor in this thesis. It is the first proposal to realize robust top-rank learning. Moreover, the robust top-rank learning framework with representation learning is combined in an end-to-end way. Experiments results demonstrate the robustness of the robust top-rank learning algorithms for outliers on the artificial datasets and medical image datasets. Benefits from the robustness to outliers, the proposed method obtains more reliable positive samples (higher Pos@Top) in the test set than the comparison

methods on medical image diagnosis tasks. More importantly, the stability of the learned ranking function was experimentally proved.

6.2 Future Work

I hope to further analyze top-rank learning with neural networks and improve its performance on different applications for future work. Several potential research directions are listed as follows.

- The effect of mini-batch for the training of top-rank learning with neural networks should be observed. As discussed in Section 4.3.2, TopRank CNN has strict requirements for each mini-batch. In future work, the effect of different sizes within each minibatch to optimize of TopRank CNN will be analyzed.
- Applying top-rank learning with neural networks to multi-class problems can be considered. In previous researches, the superiority and usefulness of Top-rank learning with neural networks on binary classes tasks have been proved. However, this characteristic limits its ability on multi-class applications. Therefore, applying top-rank learning to multi-class problems by learning different top-ranked negative samples within different classes will be considered.
- Further analysis of the characteristic of robust top-rank learning will be given. In previous researches, top-rank learning was very sensitive to the hyperparameters initialization. In the next step, the performance of top-rank learning with different initial features will be analyzed. In addition, robust top-rank learning will be utilized in other datasets and its performance will be observed.
- Improving the objective function of top-rank learning for its robustness to outliers will be considered. The current robust top-rank learning achieved good performance on weakening the effect of an outlier. Still, it relies on the additional hyperparameter, which increases the time and computational cost for

their tuning. In the future, this thesis will consider improving the objective function of top-rank learning by adding some constraints during the optimization and maintaining its robustness to outliers.

Bibliography

- [1] A. He, T. Li, N. Li, K. Wang, and H. Fu, “Cabnet: Category attention block for imbalanced diabetic retinopathy grading,” *IEEE Trans. Medical Imaging*, vol. 40, no. 1, pp. 143–153, 2021.
- [2] T. Liu, “Learning to rank for information retrieval,” *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, 2009.
- [3] T. Joachims, “Optimizing search engines using clickthrough data,” in *Proceedings of the SIGKDD*, 2002, pp. 133–142.
- [4] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, “Learning to rank using gradient descent,” in *Proceedings of the ICML*, 2005, pp. 89–96.
- [5] H. Li, “Learning to rank for information retrieval and natural language processing,” *Synthesis Lectures on Human Language Technologies*, vol. 4, no. 1, pp. 1–113, 2011.
- [6] S. Agarwal and P. Niyogi, “Stability and generalization of bipartite ranking algorithms,” in *Proceedings of the COLT*, 2005, pp. 32–47.
- [7] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, “Generalization bounds for the area under the ROC curve,” *J. Mach. Learn. Res.*, vol. 6, pp. 393–425, 2005.
- [8] A. K. Menon and R. C. Williamson, “Bipartite ranking: a risk-theoretic perspective,” *J. Mach. Learn. Res.*, vol. 17, pp. 1–102, 2016.
- [9] M. Alsulmi and B. Carterette, “Improving medical search tasks using learning to rank,” in *CIBCB*, 2018, pp. 1–8.
- [10] K. Uematsu and Y. Lee, “On theoretically optimal ranking functions in bipartite ranking,” *J. Am. Stat. Assoc.*, vol. 112, no. 519, pp. 1311–1322, 2017.
- [11] J. Hanley and B. McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

- [12] C. Cortes and M. Mohri, “AUC optimization vs. error rate minimization,” in *Proceedings of the NIPS*, 2003, pp. 313–320.
- [13] T. Joachims, “A support vector method for multivariate performance measures,” in *Proceedings of the ICML*, 2005, pp. 377–384.
- [14] H. Narasimhan and S. Agarwal, “A structural SVM based approach for optimizing partial auc,” in *Proceedings of the ICML*, 2013, pp. 516–524.
- [15] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, “A support vector method for optimizing average precision,” in *Proceedings of the SIGIR*, 2007, pp. 271–278.
- [16] S. Cl  men  on and N. Vayatis, “Ranking the best instances,” *J. Mach. Learn. Res.*, vol. 8, pp. 2671–2699, 2007.
- [17] S. P. Boyd, C. Cortes, M. Mohri, and A. Radovanovic, “Accuracy at the top,” in *Proceedings of the NIPS*, 2012, pp. 962–970.
- [18] N. Li, R. Jin, and Z. Zhou, “Top rank optimization in linear time,” in *Proceedings of the NIPS*, 2014, pp. 1502–1510.
- [19] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer, “An efficient boosting algorithm for combining preferences,” in *Proceedings of ICML*, 1998, pp. 170–178.
- [20] M. R. Amini, T. V. Truong, and C. Goutte, “A boosting algorithm for learning bipartite ranking functions with partially labeled data,” in *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*, 2008, pp. 99–106.
- [21] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013.
- [22] Y. Cao, J. Xu, T. Liu, H. Li, Y. Huang, and H. Hon, “Adapting ranking SVM to document retrieval,” in *Proceedings of the SIGIR*, 2006, pp. 186–193.
- [23] C. Lee and C. Lin, “Large-scale linear ranksvm,” *Neural Comput.*, vol. 26, no. 4, pp. 781–817, 2014.
- [24] R. Cruz, K. Fernandes, J. S. Cardoso, and J. F. P. da Costa, “Tackling class imbalance with ranking,” in *Proceedings of the IJCNN*, 2016, pp. 2182–2187.
- [25] Y. Zheng, Y. Zheng, W. Ohyama, D. Suehiro, and S. Uchida, “Ranksvm for offline signature verification,” in *Proceedings of the ICDAR*, 2019, pp. 928–933.
- [26] A. Rakotomamonjy, “Optimizing area under roc curve with svms,” in *Proceedings of ROCAI*, 2004, pp. 71–80.

- [27] R. Cruz, K. Fernandes, J. F. P. da Costa, M. Pérez-Ortiz, and J. S. Cardoso, “Combining ranking with traditional methods for ordinal class imbalance,” in *Proceedings of the IWANN*, 2017, pp. 538–548.
- [28] S. Srihari, A. Xu, and M. Kalera, “Learning Strategies and Classification Methods for Off-line Signature Verification,” in *ICFHR*, 2004.
- [29] L. Hafemann, R. Sabourin, and L. Oliveira, “Analyzing Features Learned for Offline Signature Verification Using Deep CNNs,” in *ICPR*, 2016.
- [30] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Learning features for offline handwritten signature verification using deep convolutional neural networks,” *Pattern Recognit.*, vol. 70, pp. 163–176, 2017.
- [31] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [32] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [33] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: identifying density-based local outliers,” in *Proceedings of the ICMD*, 2000, pp. 93–104.
- [34] O. Chapelle, Y. Chang, and T.-Y. Liu, “Future directions in learning to rank,” in *Proceedings of the Learning to Rank Challenge*, 2011, pp. 91–100.
- [35] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li, “Learning to Rank: From Pairwise Approach to Listwise Approach,” in *ICML*, 2007.
- [36] D. Cossock and T. Zhang, “Statistical analysis of bayes optimal subset ranking,” *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5140–5154, 2008.
- [37] R. Herbrich, “Large Margin Rank Boundaries for Ordinal Regression,” *Advances in Large-Margin Classifiers*, 2000.
- [38] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, “Learning to rank using gradient descent,” in *Proceedings of the ICML*, 2005, pp. 89–96.
- [39] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, “Listwise approach to learning to rank: theory and algorithm,” in *Proceedings of ICML*, 2008, pp. 1192–1199.
- [40] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, “An efficient boosting algorithm for combining preferences,” *Journal of Mach. Learn. Res.*, vol. 4, no. Nov, pp. 933–969, 2003.

- [41] S. Cl  men  on, G. Lugosi, and N. Vayatis, “Ranking and empirical minimization of u-statistics,” *The Annals of Statistics*, vol. 36, no. 2, pp. 844–874, 2008.
- [42] Q. Le and A. Smola, “Direct optimization of ranking measures,” *arXiv preprint arXiv:0704.3359*, 2007.
- [43] M. Xu, Y. Li, and Z. Zhou, “Multi-label learning with PRO loss,” in *Proceedings of the AAAI*, 2013.
- [44] H. Valizadegan, R. Jin, R. Zhang, and J. Mao, “Learning to rank by optimizing ndcg measure,” in *Proceedings of the NIPS*, 2009, pp. 1883–1891.
- [45] K. Wu and J. Liu, “Multi-objective evolutionary top rank optimization with pareto ensemble,” in *Proceedings of SSCI*, 2020, pp. 624–630.
- [46] C. Rudin, “The p-norm push: a simple convex ranking algorithm that concentrates at the top of the list,” *J. Mach. Learn. Res.*, vol. 10, pp. 2233–2271, 2009.
- [47] S. Agarwal, “The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list,” in *Proceedings of the SDM*, 2011, pp. 839–850.
- [48] S. P. Boyd, C. Cortes, M. Mohri, and A. Radovanovic, “Accuracy at the top,” in *Proceedings of the NIPS*, 2012, pp. 962–970.
- [49] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Offline handwritten signature verification—literature review,” in *Proceedings of IPTA*, 2017, pp. 1–8.
- [50] G. L. L. Hafemann, L. Oliveira, and R. Sabourin, “Fixed-sized Representation Learning from Offline Handwritten Signatures of Different Sizes,” *IJDAR*, 2018.
- [51] S. M. Sam, K. Kamardin, N. N. A. Sjarif, N. Mohamed *et al.*, “Offline signature verification using deep learning convolutional neural network (cnn) architectures googlenet inception-v1 and inception-v3,” *Procedia Computer Science*, vol. 161, pp. 475–483, 2019.
- [52] Y. Zheng, Y. Zheng, W. Ohyama, D. Suehiro, and S. Uchida, “Ranksvm for offline signature verification,” in *Proceedings of ICDAR*, 2019, pp. 928–933.
- [53] E. Alajrami, B. A. Ashqar, B. S. Abu-Nasser, A. J. Khalil, M. M. Musleh, A. M. Barhoom, and S. S. Abu-Naser, “Handwritten signature verification using deep learning,” *Proceedings of IJAMR*, vol. 3, no. 12, 2020.
- [54] Y. Zheng, B. K. Iwana, M. I. Malik, S. Ahmed, W. Ohyama, and S. Uchida, “Learning the micro deformations by max-pooling for offline signature verification,” *Pattern Recognit.*, vol. 118, p. 108008, 2021.

- [55] M. I. Malik and M. Liwicki, "From terminology to evaluation: Performance assessment of automatic signature verification systems," in *Proceedings of ICFHR*, 2012, pp. 613–618.
- [56] L. Liu, L. Huang, F. Yin, and Y. Chen, "Offline signature verification using a region based deep metric learning network," *Pattern Recognit.*, vol. 118, p. 108009, 2021.
- [57] J. Hu and Y. Chen, "Offline signature verification using real adaboost classifier combination of pseudo-dynamic features," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1345–1349.
- [58] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin, and S. L. Fernandes, "A framework for offline signature verification system: Best features selection approach," *Pattern Recognit. Lett.*, vol. 139, pp. 50–59, 2020.
- [59] M. Kanawade and S. Katariya, "Review of offline signature verification and recognition system," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 7, pp. 659–662, 2013.
- [60] M. A. Ferrer, J. F. Vargas, A. Morales, and A. Ordonez, "Robustness of offline signature verification based on gray level features," *IEEE Trans. Inf. Forensic Secur.*, vol. 7, no. 3, pp. 966–977, 2012.
- [61] V. L. Souza, A. L. Oliveira, R. M. Cruz, and R. Sabourin, "A white-box analysis on the writer-independent dichotomy transformation applied to offline handwritten signature verification," *Expert Syst. Appl.*, vol. 154, p. 113397, 2020.
- [62] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," in *ECML*, 2004.
- [63] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT press, 2018.
- [64] W. Elisseeff, "Kernel methods for Multi-labelled Classification and Categorical Regression Problems," *NIPS*, 2002.
- [65] R. Cruz, K. Fernandes, J. Costa, M. Ortiz, and J. Cardoso, "Ordinal Class Imbalance with Ranking," in *IbPRIA*, 2017.
- [66] M. Ferrer, M. Diaz, C. Carmona-Duarte, and A. Morales, "A Behavioral Handwriting Model for Static and Dynamic Signature Synthesis," *TPAMI*, 2017.
- [67] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. P. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the CIKM*, 2013, pp. 2333–2338.

- [68] A. Severyn and A. Moschitti, “Learning to rank short text pairs with convolutional deep neural networks,” in *Proceedings of the SIGIR*, 2015, pp. 373–382.
- [69] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, “A deep relevance matching model for ad-hoc retrieval,” in *Proceedings of the CIKM*, 2016, pp. 55–64.
- [70] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” in *Proceedings of the CVPR*, 2014, pp. 1386–1393.
- [71] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the CVPR*, 2015, pp. 815–823.
- [72] K. Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” in *Proceedings of the NIPS*, 2016, pp. 1857–1865.
- [73] S. Chen, C. Guo, and J. Lai, “Deep ranking for person re-identification via joint representation learning,” *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, 2016.
- [74] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, “Learning semantic representations using convolutional neural networks for web search,” in *Proceedings of the WWW*, 2014, pp. 373–374.
- [75] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, and X. Cheng, “Deeprank: A new deep architecture for relevance ranking in information retrieval,” in *Proceedings of the CIKM*, 2017, pp. 257–266.
- [76] D. Di, S. Li, J. Zhang, and Y. Gao, “Ranking-based survival prediction on histopathological whole-slide images,” in *Proceedings of MICCAI*, 2020, pp. 428–438.
- [77] J. Lyu, S. H. Ling, S. Banerjee, J. Zheng, K.-L. Lai, D. Yang, Y.-P. Zheng, X. Bi, S. Su, and U. Chamoli, “Ultrasound volume projection image quality selection by ranking from convolutional ranknet,” *Computerized Medical Imaging and Graphics*, vol. 89, p. 101847, 2021.
- [78] D. Song, W. Liu, R. Ji, D. A. Meyer, and J. R. Smith, “Top rank supervised binary coding for visual search,” in *Proceedings of the ICCV*, 2015, pp. 1922–1930.
- [79] Y. Geng, R.-Z. Liang, W. Li, J. Wang, G. Liang, C. Xu, and J. Y. Wang, “Learning convolutional neural network to maximize pos@ top performance measure,” in *Proceedings of the ESANN*, 2017, pp. 589–594.

- [80] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.
- [81] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," URL <http://www.cs.toronto.edu/kriz/cifar.html>, vol. 8, 2010.
- [82] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the ICCV*, 2017, pp. 2999–3007.
- [83] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [84] K. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, M. Sallam, K. Chang, and K. Woods, "The digital database for screening mammography," in *Proceedings of the IWDM*, 1996, p. 27.
- [85] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, p. 170177, 2017.
- [86] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in *Proceedings of the ICML*, P. Langley, Ed., 2000, pp. 255–262.
- [87] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLOS ONE*, vol. 11, no. 4, p. e0152173, 2016.
- [88] J. Kim and C. D. Scott, "Robust kernel density estimation," *Journal of Mach. Learn. Res.*, vol. 13, pp. 2529–2565, 2012.
- [89] E. Eskin, A. Arnold, M. J. Prerau, L. Portnoy, and S. J. Stolfo, "A geometric framework for unsupervised anomaly detection," in *Applications of Data Mining in Computer Security*. Springer, 2002, pp. 77–101.
- [90] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proceedings of the ICCV*, 2015, pp. 1511–1519.
- [91] H. Xu, C. Caramanis, and S. Mannor, "Outlier-robust PCA: the high-dimensional case," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 546–572, 2013.
- [92] G. Truong, H. Le, D. Suter, E. Zhang, and S. Z. Gilani, "Unsupervised learning for robust fitting: A reinforcement learning approach," in *Proceedings of CVPR*, 2021, pp. 10 348–10 357.

- [93] X. Chen, J. Weng, W. Luo, W. Lu, H. Wu, J. Xu, and Q. Tian, “Sample balancing for deep learning-based visual recognition,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 3962–3976, 2020.
- [94] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.
- [95] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *Proceedings of the CVPR*, 2014, pp. 1717–1724.
- [96] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, “Feedback on a publicly distributed database: the messidor database,” *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, 2014.
- [97] Y. Zheng, Y. Zheng, W. Ohyama, D. Suehiro, and S. Uchida, “Ranksvm for offline signature verification,” in *Proceedings of the ICDAR*, 2019, pp. 928–933.
- [98] Y. Zheng, Y. Zheng, D. Suehiro, and S. Uchida, “Top-rank convolutional neural network and its application to medical image-based diagnosis,” *Pattern Recognit.*, vol. 120, p. 108138, 2021.