

Many-to-One Distribution Learning and K -Nearest Neighbor Smoothing for Thoracic Disease Identification

Yi Zhou^{1*}, Lei Huang², Tianfei Zhou³, Ling Shao⁴

¹School of Computer Science and Engineering, Southeast University, Nanjing, China

²State Key Laboratory of Software Development Environment, Beihang University, Beijing, China

³Computer Vision Laboratory, ETH Zurich, Switzerland

⁴Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

yi.zhou@ieee.org

Abstract

Chest X-rays are an important and accessible clinical imaging tool for the detection of many thoracic diseases. Over the past decade, deep learning, with a focus on the convolutional neural network (CNN), has become the most powerful computer-aided diagnosis technology for improving disease identification performance. However, training an effective and robust deep CNN usually requires a large amount of data with high annotation quality. For chest X-ray imaging, annotating large-scale data requires professional domain knowledge and is time-consuming. Thus, existing public chest X-ray datasets usually adopt language pattern based methods to automatically mine labels from reports. However, this results in label uncertainty and inconsistency. In this paper, we propose many-to-one distribution learning (MODL) and K -nearest neighbor smoothing (KNNS) methods from two perspectives to improve a single model's disease identification performance, rather than focusing on an ensemble of models. MODL integrates multiple models to obtain a soft label distribution for optimizing the single target model, which can reduce the effects of original label uncertainty. Moreover, KNNS aims to enhance the robustness of the target model to provide consistent predictions on images with similar medical findings. Extensive experiments on the public NIH Chest X-ray and CheXpert datasets show that our model achieves consistent improvements over the state-of-the-art methods.

Introduction

Chest X-rays are one of the most common radiology exams. Hundreds of millions of such images are acquired in hospitals and clinics all over the world, for identifying a wide range of diseases. Advances in deep learning present a powerful means of developing automated systems to help radiologists interpret chest X-ray images. Recent efforts have shown promise in improving the identification of different chest lesions (Rajpurkar et al. 2017; Allaouzi and Ahmed 2019), localization of their corresponding positions and sizes (Li et al. 2018b; Liu et al. 2019), and generation of human-readable diagnostic reports (Jing, Xie, and Xing 2018; Li et al. 2018a; Zhang et al. 2020). However, there exist many drawbacks to chest X-rays, such as the lack of

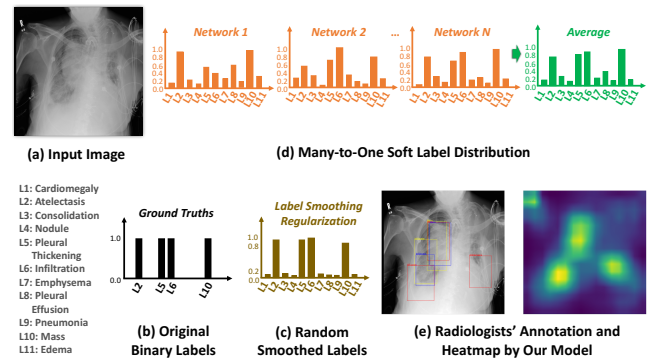


Figure 1: (a) An input image. (b) The ground-truth binary labels automatically extracted from reports. (c) Random smoothed labels proposed by (Pham et al. 2019). (d) Our many-to-one soft label distribution. (e) The bounding boxes of three different colors are annotated by three radiologists, separately, in the left image. The right image is the heatmap, showing abnormal regions, obtained by our method.

3D information, inconsistent visual characteristics of various lesions, and label uncertainty, which still make automated chest X-ray diagnosis a very challenging task.

For chest X-ray images, large anonymized public datasets are available to researchers, and have facilitated the development of deep learning models. However, providing accurate clinical labels for the very large image sets required for training deep models is difficult and time-consuming. Most datasets (Wang et al. 2017; Irvin et al. 2019) have applied rule-based natural language processing (NLP) to radiology reports to automatically mine the labels, but this often leads to inconsistencies and noise for training. Although previous efforts have tried to address this problem, for example, by converting all the uncertain labels into positive or negative (Irvin et al. 2019), or applying label smoothing regularization (Pham et al. 2019), these methods are not quite effective and only bring a slight improvement. Moreover, most efforts focus on model ensembles when pursuing a further increase of accuracy, rather than enhancing a single model. Combining multiple models, however, will largely increase the computational costs and complexities. Therefore, in this

*Corresponding author: Yi Zhou

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

paper, we aim to address these two challenges by exploring label distributions to reduce the effects of uncertainty and noise, and enhancing a single model’s ability rather than using a model ensemble. Figure 1 illustrates our motivation and solutions.

To train deep learning models for identifying thoracic diseases, multi-label supervision indicating different lesions is usually adopted. The optimization of such models is treated as an image classification problem that builds a mapping from the instances to the labels via a Softmax cross-entropy or binary cross-entropy loss for each label. However, binary labels (positive or negative) cannot usually describe in detail the severity of a certain disease in an image. As shown in Figure 1 (e), even radiologists are unable to always provide consistent annotations. With the label uncertainty and error further introduced by automatic label extraction, we believe using soft labels of probabilities will contribute to more reasonable training. Thus, in addition to multi-label classification, we propose to optimize the Kullback-Leibler divergence between the ground-truth soft label distribution and the predicted one. Moreover, to obtain a single model with competitive performance to model ensembles, we compute K nearest neighbors using the soft label distributions integrated from multiple models, and then constrain a single target model to provide similar predictions to these scans, which can enhance its robustness. Our main contributions are highlighted as follows:

1. A many-to-one distribution learning (MODL) method is proposed to exploit multiple reference models and integrate their predictions as a soft label distribution to optimize the target network, which can reduce the effects of label uncertainty and ambiguity.
2. To enhance the robustness of a target model for providing consistent predictions on images with similar medical findings, we introduce a K -nearest neighbor smoothing (KNNS) module as an auxiliary loss for optimization. The local similarity score is adopted to enhance the learning of the neighbor smoothing.
3. Extensive experiments are conducted to demonstrate the effectiveness of the proposed MODL and KNNS. Consistent improvements by our method are achieved on both the NIH Chest X-ray and CheXpert datasets for thoracic disease identification, compared to state-of-the-art models. Another big advantage of our method is that our single target model does not require additional computational resources during the testing phase.

Related Work

Thoracic Disease Diagnosis

Recently, deep convolutional neural networks (CNNs) have been extensively investigated for thoracic disease diagnosis (Litjens et al. 2017; Shin et al. 2016; Zhou et al. 2021; Fan et al. 2020). The common tasks can be categorized into multi-disease identification (Kumar, Grewal, and Srivastava 2018; Rajpurkar et al. 2017; Zhou, Li, and Wang 2018; Baltruschat et al. 2019), localization (Pesce et al. 2019; Yao et al. 2018; Liu et al. 2019), and automatic report generation (Wang et al. 2018; Jing, Xie, and Xing 2018; Li et al.

2018a). To facilitate research into thoracic disease diagnosis, various public chest X-ray datasets (Wang et al. 2017; Irvin et al. 2019; Johnson et al. 2019) have been released. Several single- or multi-CNN models (Rajpurkar et al. 2017) have been developed for disease classification, while weakly supervised mechanisms using limited bounding-box annotations are typically used for localization. Li *et al.* (Li et al. 2018b) proposed a patch slicing layer to resize CNN features by max-pooling or bilinear interpolation and applied a fully convolutional recognition network on these regions, to improve identification and localization, simultaneously. Moreover, a contrast induced attention network was introduced by (Liu et al. 2019) to exploit the highly structured property of chest X-ray images for localizing diseases. In addition, the automatic generation of medical image reports has also been investigated; for example, (Jing, Xie, and Xing 2018) exploited a co-attention mechanism to identify abnormal regions and used a hierarchical LSTM to generate sentences. However, while several previous works have mentioned issues with label uncertainty and inconsistency of annotation, but none have provided an effective solution.

Reducing Label Inconsistency and Ambiguity

Label inconsistency, uncertainty, and ambiguity are common problems that affect the training of machine learning models in numerous areas. Several strategies, such as confident learning (CL) (Northcutt, Jiang, and Chuang 2019), label distribution learning (LDL) (Geng 2016; Gao et al. 2017; Chen et al. 2020), label enhancement (LE) (Xu, Lv, and Geng 2019; Xu, Tao, and Geng 2018), and active learning (AL) (Wu et al. 2018), have been proposed in recent years to mitigate these problems. CL is based on the principles of pruning noisy data, counting to estimate noise, and ranking examples to train with confidence. It can directly estimate the joint distribution of noisy and true labels, and find the label errors. LDL treats the labels representing the degree to which each label describes the instance. It is a more general learning framework of which single-label and multi-label classification are special cases. LDL has been adopted in many applications, including facial landmark detection (Su and Geng 2019) and age estimation (Hou et al. 2017). LE methods aim to solve the unavailability of label distributions. For example, a partial label learning method was presented by (Xu, Lv, and Geng 2019), in which label distributions are recovered by leveraging the topological information of the feature space. Moreover, AL adopts a small set of clean data, and decreases the noise of the training set by re-labeling other uncertain samples.

Proposed Methods

Problem Formulation

First of all, the main notations used in this paper are defined as follows. Given the i -th input X-ray image x_i and its corresponding binary ground truth labels y_i^j , where j denotes the j -th disease label, we aim to optimize a single target network f_T for thoracic disease identification. Due to the uncertainty of the original binary labels, we train multiple reference models $f_n, n \in \{1, 2, \dots, N\}$, where N is

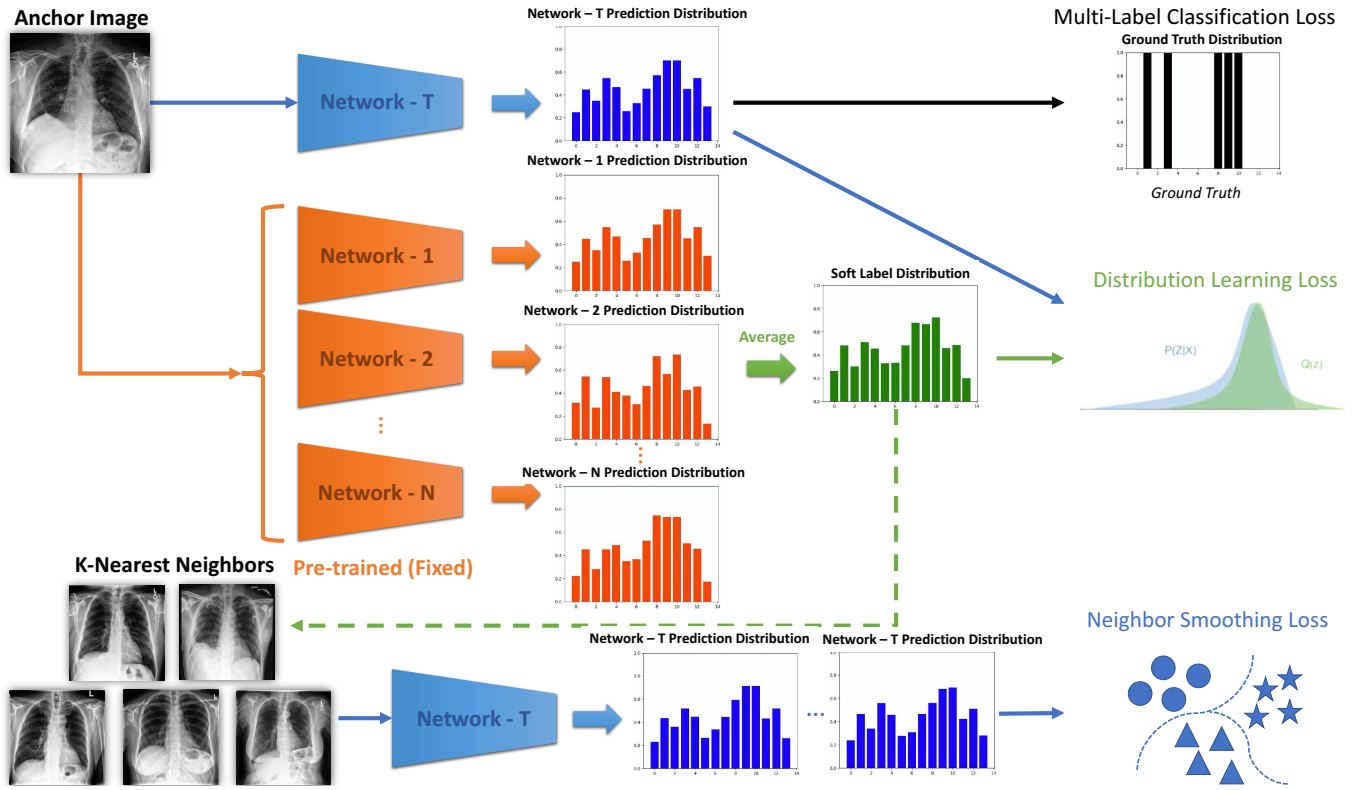


Figure 2: Overview of the proposed method. Network-T is the single target network, which we aim to optimize. N reference networks are pre-trained to compute soft label distributions for helping to optimize Network-T with a distribution learning loss, which can reduce the uncertainty and ambiguity of the original binary labels. Moreover, K nearest neighbors are computed based on the soft label distribution, and then used to constrain the Network-T to provide consistent predictions using an auxiliary neighbor smoothing loss.

the number of networks, and integrate their pseudo probability predictions as the input image’s soft label distribution $l_i = (l_{x_i}^{y_1}, l_{x_i}^{y_2}, \dots, l_{x_i}^{y_C})$, where C is the number of label classes. Thus, in addition to the standard multi-label classification loss \mathcal{L}_{cls} , we propose a label distribution learning loss \mathcal{L}_{distr} to better optimize the target f_T for reducing the effects of the label uncertainty. Moreover, we indicate the input image as an anchor, and compute its K nearest neighbors $x_i^k, k \in \{1, 2, \dots, K\}$ based on the averaged soft label predictions. An auxiliary neighbor smoothing loss \mathcal{L}_{neigh} is introduced to enhance the robustness of the target f_T . Therefore, the target model can be optimized by minimizing the overall loss function as:

$$\mathcal{L} = \sum_{i,j} [\mathcal{L}_{cls}(y_i^j, f_T(x_i|\theta)) + \lambda \mathcal{L}_{distr}(l_i, f_T(x_i|\theta))] \quad (1) \\ + \gamma \sum_k \mathcal{L}_{neigh}(f_T(x_i|\theta), f_T(x_i^k|\theta)),$$

where θ is the parameter vector, and λ and γ balance the weights of different losses. An overview of our proposed method is illustrated in Figure 2.

Basic Multi-Label Classification and Label Smoothing Regularization

The basic thoracic disease identification framework used in most previous works (Wang et al. 2017; Rajpurkar et al. 2017), usually adopts a classic deep CNN (i.e. backbone) and optimizes it with a binary cross-entropy (BCE) loss for multi-label classification. DenseNet-121 (Huang et al. 2017) is the most commonly used backbone for this task, since it improves the flow of information and gradients through the network, making the optimization of deep networks tractable. Therefore, we select DenseNet-121 as our target network f_T for final evaluation, and use the BCE loss as \mathcal{L}_{cls} for training.

To deal with the label uncertainty caused by the ground-truth labels not being annotated by human experts for images provided in the training dataset, several preliminary studies have been conducted. For example, for the large-scale public CheXpert (Irvin et al. 2019) dataset, uncertain labels are often extracted due to the unavoidable ambiguities in radiology reports. These labels can all be *ignored* (U-Ignore), converted to *positive* labels (U-Ones), or converted to *negative* labels (U-Zeros), proposed in (Irvin et al. 2019). Moreover, in (Pham et al. 2019), label smoothing regularization (LSR) was applied to better handle these uncer-

tain labels, with the main goal of preventing the model from training with overconfident labels that might contain misclassified data. Specifically, in the U-Ones setting, the uncertainty labels are mapped to a random number close to one, subject to a uniform distribution $\mathcal{U}(a, b)$, where a and b are the hyper-parameters. In contrast, in the U-Zeros setting, a similar process is used to convert the uncertainty labels to a random number close to zero. Although these solutions provide some improvements, they are based on too much on brute-force to address the mislabeled data reasonably.

Many-to-One Distribution Learning

Incorporating the opinions of multiple radiologists is usually more reliable than only adopting a single diagnosis. Similarly, although the target network f_T based on the single backbone DenseNet-121 can achieve satisfactory results, model ensemble of several network architectures usually further makes a big improvement on the mean accuracy. According to many observations from extensive experiments, different network architectures, and even different epoch checkpoints for one network architecture, provide inconsistent predictions and have their own strengths on different diseases. Thus, integrating probability predictions from multiple models can reduce the effects of the uncertainty and ambiguity from the original binary labels. In this paper, we propose a many-to-one distribution learning method to optimize a single target network by taking the essence and discarding the dregs of multiple networks, which can also avoid the extra computational resources introduced by a model ensemble in the testing phase.

To select a strong set of reference models, we follow (Pham et al. 2019) and adopt six state-of-the-art CNNs ($N = 6$), including DenseNet-121, DenseNet-169, DenseNet-201 (Huang et al. 2017), Inception-ResNet-v2 (Szegedy et al. 2017), Xception (Chollet 2017), and NASNetLarge (Zoph et al. 2018). Once the training is separately done for each network, the probability predictions of all the trained networks are simply averaged, and can then be viewed as the soft label distribution. By getting an average, those evident lesions with high probabilities predicted by most of the reference networks still get positive-toward values, and healthy scans still get negative-toward values. However, those lesions getting inconsistent predictions which have ambiguities will be smoothed to reduce the effects of uncertainty. Our label distribution smoothing tends to be a major voting manner, which is more reasonable than the random LSR method. Therefore, this soft label distribution for each image should be closer to the correct labels to indicate the ground truths. We use the Kullback-Leibler (KL) divergence for the distribution learning loss \mathcal{L}_{distri} and minimize it to optimize the target network f_T :

$$\mathcal{L}_{distri} = KL\{l_i || f_T(x_i|\theta)\} = \sum_i l_i \log\left(\frac{l_i}{f_T(x_i|\theta)}\right). \quad (2)$$

Please note that all the reference networks f_n are pre-trained in advance. Their network parameters are fixed when training the target network f_T . Thus, the soft label distribu-

tion l_i is fixed to compute the KL divergence as the relative entropy of l_i with respect to $f_T(x_i|\theta)$.

K-Nearest Neighbor Smoothing

Our second important concern is how to make the target network as robust and competitive as the model ensemble. According to the smoothness assumption (Chapelle, Scholkopf, and Zien 2009), we assume that if two samples are similar in the averaged label distribution space, they should also be similar in the label space predicted by the target model. Therefore, before training the target network, the K nearest neighbors of an image (marked as the anchor image) are pre-computed based on the fixed soft label distribution.

To make the disease predictions of the anchor image $f_T(x_i|\theta)$ as close as the predictions of its corresponding k -th neighbor $f_T(x_i^k|\theta)$, we generate a pool that contains K neighbor images of each anchor during the training, for enhancing local similarity. Let k denote a sample in the pool. Inspired by (Xu, Tao, and Geng 2018), we specify the local similarity s_i^k which is defined as:

$$s_i^k = \begin{cases} \exp(-\frac{\|l_i - l_i^k\|^2}{2\sigma^2}) & \text{if } k \in \{1, \dots, K\} \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where σ denotes the width parameter for similarity calculation, which is fixed as 1 in our settings. Thus, the more similar l_i and l_i^k are, the higher s_i^k will be, which means $f_T(x_i|\theta)$ and $f_T(x_i^k|\theta)$ should be closer. We propose a neighbor smoothing loss \mathcal{L}_{neigh} as follows:

$$\mathcal{L}_{neigh} = \sum_{i,k} s_i^k f_T(x_i^k|\theta) \log \frac{f_T(x_i^k|\theta)}{f_T(x_i|\theta)}. \quad (4)$$

By optimizing \mathcal{L}_{neigh} , the target model f_T is constrained to give similar predictions of neighboring images which have similar abnormal findings. This enhances the robustness of f_T for more consistent identification performance.

Implementation Details

The training procedure of our proposed method consists of two steps. In the first step, all the reference models based on different backbones are pre-trained using the basic multi-label classification loss \mathcal{L}_{cls} . All the images are resized into a standard size 256×256 and randomly cropped to 224×224 patches as inputs. The Adam optimizer is adopted with an initial learning rate of 0.001 and default parameters $\beta_1 = 0.9, \beta_2 = 0.999$. The network parameters are initialized with a model pre-trained on ImageNet (Deng et al. 2009). The mini-batch size is set to 32 for training over 10 epochs and the learning rate is reduced by a factor of 3 after every two epochs. Once all the reference models are well trained, the soft label distribution of each image is computed and used for training in the second step.

In the second step of the training phase, in addition to \mathcal{L}_{cls} , we add the proposed MODL loss \mathcal{L}_{distri} and KNNs loss \mathcal{L}_{neigh} for optimizing the single target model. For each image in a training batch (we marked it as an anchor), its

Methods	Atel.	Card.	Effu.	Infi.	Mass	Nod.	Pne.1	Pne.2	Cos.	Ede.	Emp.	Fibr.	P.T.	Hern	Mean
(Wang et al. 2017)	70.03	81.00	75.85	66.14	69.33	66.87	65.80	79.93	70.32	80.52	83.30	78.59	68.35	87.17	74.51
(Yao et al. 2018)	73.30	85.80	80.60	67.50	72.70	77.80	69.00	80.50	71.70	80.60	84.20	75.70	72.40	82.40	76.73
(Yao et al. 2017)	77.20	90.40	85.90	69.50	79.20	71.70	71.30	84.10	78.80	88.20	82.90	76.70	76.50	91.40	80.30
(Li et al. 2018b)	80.00	81.00	87.00	70.00	83.00	75.00	67.00	87.00	80.00	88.00	91.00	78.00	79.00	77.00	80.21
(Rajpurkar et al. 2017)	80.94	92.48	86.38	73.45	86.76	78.02	76.80	88.87	79.01	88.78	93.71	80.47	80.62	91.64	84.13
Baseline (B)	82.42	89.80	89.68	68.46	86.84	80.34	75.31	85.34	80.37	91.62	90.82	84.60	80.64	93.97	84.30
B+MODL	83.91	91.66	91.58	72.84	88.68	81.65	76.73	89.32	82.19	93.72	93.83	86.51	82.92	95.69	86.52
B+KNNS	84.44	91.73	91.62	72.73	88.99	82.12	76.97	89.52	82.11	93.75	93.81	86.78	83.74	95.77	86.72
B+MODL+KNNS	85.20	92.24	92.09	73.01	89.74	82.94	78.30	90.46	83.04	94.42	94.32	87.64	84.26	95.28	87.35

Table 1: Identification results using AUC (%) metric on the NIH Chest X-ray dataset.

Methods	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean
(Allaouzi and Ahmed 2019) LP	72.0	87.0	77.0	87.0	90.0	82.6
(Allaouzi and Ahmed 2019) BR	72.0	88.0	77.0	87.0	90.0	82.8
(Allaouzi and Ahmed 2019) CC	70.0	87.0	74.0	86.0	90.0	81.4
(Irvin et al. 2019) U-Zeros *	81.1	84.0	93.2	92.9	93.1	88.8
(Irvin et al. 2019) U-Ones *	85.8	83.2	89.9	94.1	93.4	89.3
(Pham et al. 2019) U-Zeros+CT+LSR	80.6	83.3	92.9	93.3	92.1	88.4
(Pham et al. 2019) U-Ones+CT+LSR	82.5	85.5	93.7	93.0	92.3	89.4
Baseline (B)	87.29	84.47	92.09	91.66	91.34	89.37
B+MODL	87.64	87.84	93.34	94.58	93.35	91.35
B+KNNS	89.15	88.24	92.40	94.92	92.67	91.48
B+MODL+KNNS	89.46	89.67	93.80	95.87	94.35	92.63

Table 2: Identification results using AUC (%) metric on the CheXpert dataset. * indicates that the result is obtained by the ensemble of 30 checkpoints.

K neighbor images and the corresponding local similarity scores s_i^k are stored in a pool for optimizing \mathcal{L}_{neigh} . The different loss weights λ and γ in Eq. 1 are both set to 0.1 in the experiments. All other training settings in the second step are the same as Step 1. Once the training is completed, only the target model f_T is deployed in the testing phase, without any additional computation consumption introduced by model ensemble.

Experiments and Results

Datasets

NIH Chest X-ray (Wang et al. 2017) is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients, 51,708 of which contain one or more pathologies, while the remaining images have no findings. The 14 different labels are extracted using NLP to text-mine disease classifications from the associated radiology reports, but more than 10% of the annotations are incorrect. The dataset split setting of experiments for our model follows (Rajpurkar et al. 2017).

CheXpert (Irvin et al. 2019) consists of 224,316 chest radiographs of 65,240 patients. The disease labels of the training set are labeled for the presence of 14 observations as positive, negative, uncertain, or unmentioned, which also contain inconsistency and ambiguity. However, CheXpert releases a separate validation set of 200 studies, which is annotated by three board-certified radiologists for five diseases: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. The evaluation on this set is more convincing.

Comparison with State-of-the-Arts

To evaluate the multi-disease identification task, the area-under-the-curve (AUC) of the receiver operating characteristic (ROC) is usually adopted. Table 1 provides the identification results on the NIH Chest X-ray dataset. Model (Wang et al. 2017), using a ResNet-50 backbone and weighted binary cross-entropy loss, obtains a preliminary mean AUC of 74.51% over 14 diseases. CheXnet (Rajpurkar et al. 2017) adopts DenseNet-121 and obtains a large improvement, since DenseNets improve the flow of information and gradients through the network, making the optimization of deep networks tractable. In (Yao et al. 2017), a long short-term memory (LSTM) architecture is used to model dependencies among labels, but achieves poor results. For most diseases, our method with many-to-one distribution learning and K -nearest neighbor smoothing obtains consistent increases in identification performance, compared to the best-performing method (Rajpurkar et al. 2017). The only exception is on cardiomegaly and infiltration. The mean AUC is increased by 3.22%.

The results on CheXpert are compared in Table 2. Most methods exploit ensembles of multiple models to pursue higher scores. For example, an ensemble of 30 generated checkpoints of DenseNet-121 (Irvin et al. 2019) achieves a mean AUC of 89.3%. In this subsection, we mainly focus on comparing the results achieved by single models. Binary relevance (BR), label powerset (LP), and classifier chain (CC) are explored in terms of label dependencies in (Allaouzi and Ahmed 2019), but no satisfactory performance is achieved. Conditional training (CT) and label smoothing regularization (LSR) are used to enhance the performance of a single DenseNet with a small gain achieved. In

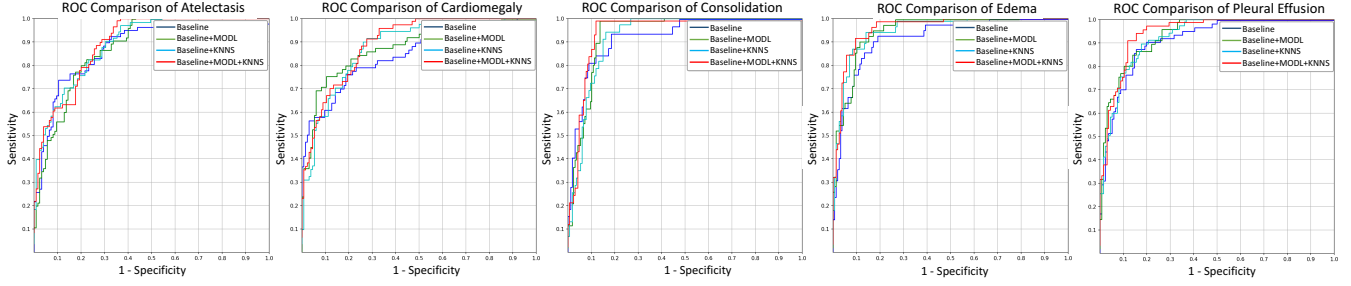


Figure 3: ROC curves of different methods for the five pathologies on the CheXpert dataset.

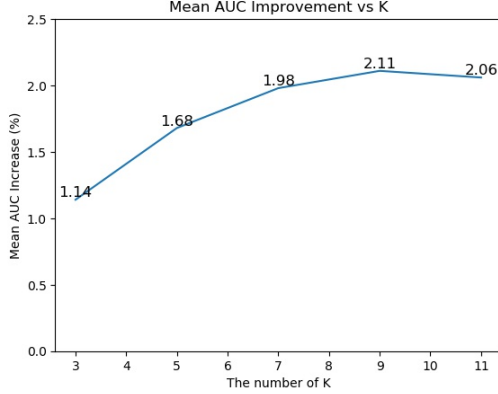


Figure 4: Comparison of the mean AUC increase based on different K in KNNS.

our implementation, the U-Ones setting is adopted to deal with those uncertainty labels when training the Baseline (**B**) (based on the backbone DenseNet-121), as well as when training the reference models in the first training stage. Our method **B**+MODL+KNNS without any ensemble operation increases the mean AUC by 3.26%.

Ablation Studies

Effectiveness of MODL To evaluate the effectiveness of the proposed many-to-one distribution learning, the soft label distribution obtained is only used to optimize the target network through the auxiliary loss $\mathcal{L}_{distrib}$. Clear improvements in mean AUC 2.22% and 1.98% are achieved on NIH Chest X-ray and CheXpert, respectively, compared to the baselines. This illustrates that the soft label distribution can represent the ground truths of a scan better than the original automatically extracted binary labels which have uncertainty and ambiguity.

Effectiveness of KNNS The K -nearest neighbor smoothing is also separately validated. An anchor image and its K neighbor images are passed forward to the target network, which is updated by the neighbor smoothing loss \mathcal{L}_{neigh} . The results show that constraining the model to provide consistent predictions for images with similar medical findings also largely improves the performance. The mean AUC is

Methods	Mean AUC	Params.
DenseNet-121 (B)	89.66	7.98 M
B +MODL+KNNS	92.63	7.98 M
Model Ensemble 1	92.19	42.13 M
Model Ensemble 2	93.96	208.33 M

Table 3: Comparison with model ensembles on the CheXpert dataset. M denotes million.

increased by 2.42% and 2.11% on NIH Chest X-ray and CheXpert, respectively. Moreover, combining $\mathcal{L}_{distrib}$ and \mathcal{L}_{neigh} can further enhance the target model without requiring any extra computational resources during the testing phase. Figure 3 provides the ROC curves of different methods for the five pathologies on the CheXpert validation set.

Investigation of the Number K of KNNS Another concern is to determine the best-performing K of KNNS. We conduct an experiment on the CheXpert dataset, where K is increased from 3 to 11 by a step of 2. The mean AUC improvement in terms of K is shown in Figure 4. We observe that configuring three neighbor images to constrain the model prediction consistency can already obtain a significant increase in mean AUC of 1.14%. Improvements of 1.68% and 1.98% are achieved when we increase K to 5 and 7, respectively. However, with the further growth of K , the increase rate becomes smaller and the method requires much higher computational memory during the training phase. Therefore, considering the trade-off between performance and computational costs, we set K as 9 in KNNS.

Comparison with Model Ensembles In addition to the mitigation of label uncertainty and enhancement of model robustness, another big advantage of our method is the zero-increase in network complexity during the testing phase, compared to model ensemble methods. To validate this point, the mean AUC and number of network parameters of some ensemble models are compared in Table 3. Model Ensemble 1 is a combination of DenseNet-121, DenseNet-169, and DenseNet-201. Model Ensemble 2 is an ensemble of the six reference models used for learning the soft label distribution in our method. As shown from the results, our method **B**+MODL+KNNS outperforms Model Ensemble 1 by 0.44% of the mean AUC, with much fewer parameters. Moreover, although Model Ensemble 2 obtains better identi-

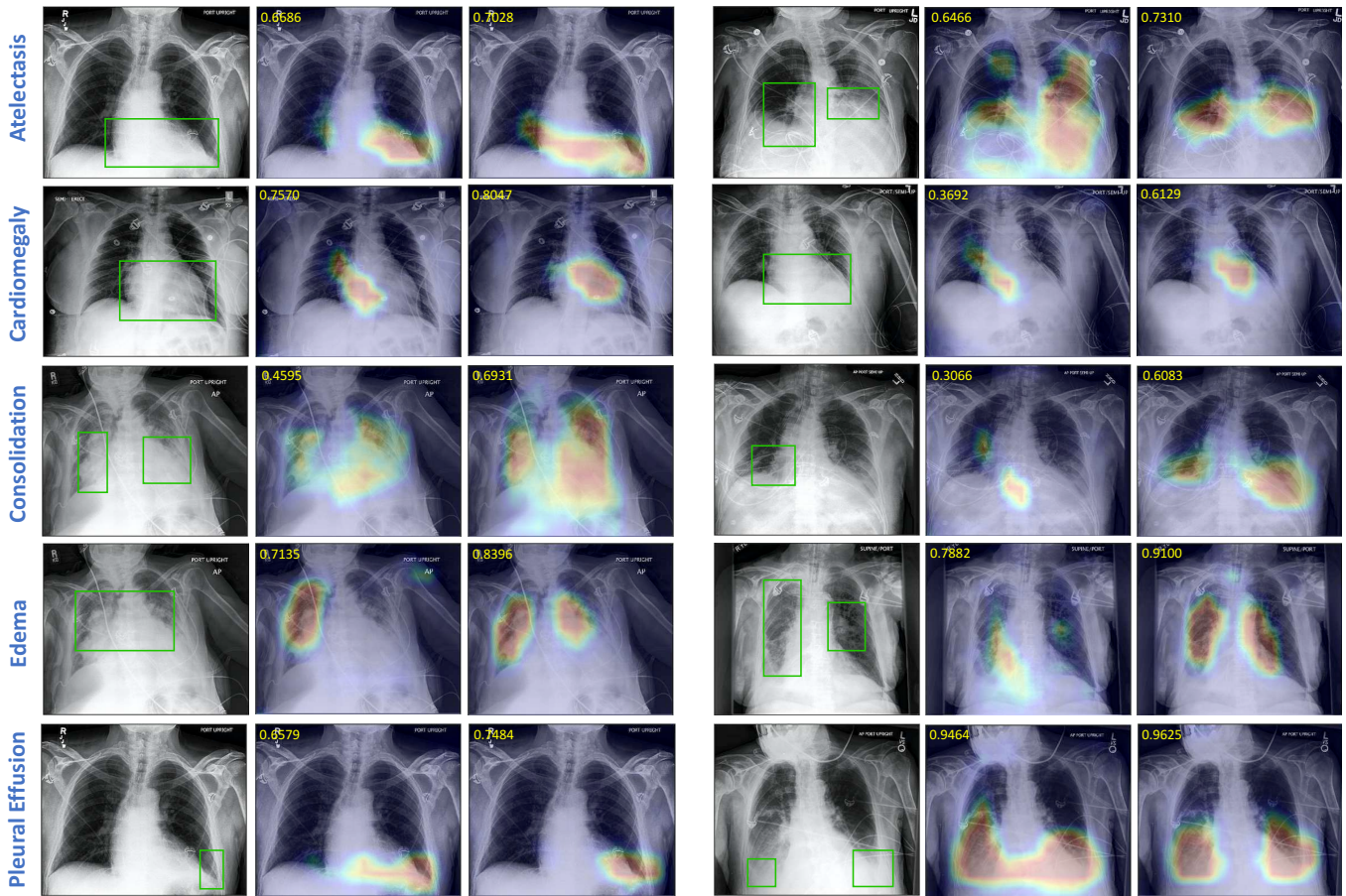


Figure 5: Heatmap visualization on the CheXpert dataset. Each row shows two cases of one disease. For each case, the left is the input image with ground truths. The middle and right are the heatmaps obtained by the Baseline (**B**) and **B+MODL+KNNS**, respectively. The value on the top-left of each heatmap is the predicted probability of the corresponding disease.

fication performance than ours, it has 26 times more parameters than our target model, making it too heavy to deploy in practical applications.

Qualitative Results

To better visualize the model performance, we adopt the probabilistic class activation map (PCAM) (Ye et al. 2020) pooling to predict the lesion heatmaps with only image-level supervision. As shown in Figure 5, we visualize the heatmaps of the five pathologies in the CheXpert dataset. All the examples do have the corresponding pathology and are labeled positive. For each case, we compare the heatmaps obtained by the Baseline (**B**) and **B+MODL+KNNS**. In addition to the improvements in identification results demonstrated before, we observe that a more satisfactory localization performance can also be achieved. For all the pathologies, the lesion regions predicted by **B+MODL+KNNS** are more precise in location and size than those predicted by the Baseline (**B**). For example, the cardiomegaly heatmaps predicted by the Baseline (**B**) usually have an abnormal region around the location of the mediastinum, while **B+MODL+KNNS** can better highlight the enlarged heart

shape region. To localize the position of pleural effusion, **B+MODL+KNNS** can better identify the left heart border, costophrenic angle and hemidiaphragm are obscured, and slight blunting of the right costophrenic angle. Moreover, as shown from the predicted probabilities in the top-left of each heatmap, **B+MODL+KNNS** identifies the positive diseases as true with higher confidence scores, compared to the Baseline (**B**).

Conclusion

In this paper, we proposed a many-to-one distribution learning and a K -nearest neighbor smoothing method to improve the performance of thoracic disease diagnosis. The methods can reduce the effects of label uncertainty and ambiguity, and also constrain the target model to provide consistent predictions on images with similar medical findings. Extensive experiments demonstrated that our method is effective and achieves significant improvements without requiring any additional computational costs in the testing phase. In our future work, we will explore visual attention methods to learn better representations of thoracic diseases for improvements.

References

- Allaouzi, I.; and Ahmed, M. B. 2019. A novel approach for multi-label chest X-ray classification of common thorax diseases. *IEEE Access* 7: 64279–64288.
- Baltruschat, I. M.; Nickisch, H.; Grass, M.; Knopp, T.; and Saalbach, A. 2019. Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific reports* 9(1): 1–10.
- Chapelle, O.; Scholkopf, B.; and Zien, A. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks* 20(3): 542–542.
- Chen, S.; Wang, J.; Chen, Y.; Shi, Z.; Geng, X.; and Rui, Y. 2020. Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition. In *CVPR*, 13984–13993. IEEE.
- Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 1251–1258. IEEE.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.
- Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging* 39(8): 2626–2637.
- Gao, B.-B.; Xing, C.; Xie, C.-W.; Wu, J.; and Geng, X. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26(6): 2825–2838.
- Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7): 1734–1748.
- Hou, P.; Geng, X.; Huo, Z.-W.; and Lv, J.-Q. 2017. Semi-supervised adaptive label distribution learning for facial age estimation. In *AAAI*.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708. IEEE.
- Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghighi, B.; Ball, R.; Shpankaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 590–597.
- Jing, B.; Xie, P.; and Xing, E. 2018. On the automatic generation of medical imaging reports. In *ACL*, 2577–2586.
- Johnson, A. E.; Pollard, T. J.; Berkowitz, S.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Kumar, P.; Grewal, M.; and Srivastava, M. M. 2018. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *International Conference Image Analysis and Recognition*, 546–552. Springer.
- Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018a. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NeurIPS*, 1530–1540.
- Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.-J.; and Fei-Fei, L. 2018b. Thoracic disease identification and localization with limited supervision. In *CVPR*, 8290–8299. IEEE.
- Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42: 60–88.
- Liu, J.; Zhao, G.; Fei, Y.; Zhang, M.; Wang, Y.; and Yu, Y. 2019. Align, Attend and Locate: Chest X-ray Diagnosis via Contrast Induced Attention Network with Limited Supervision. In *ICCV*, 10632–10641. IEEE.
- Northcutt, C. G.; Jiang, L.; and Chuang, I. L. 2019. Confident learning: Estimating uncertainty in dataset labels. *arXiv preprint arXiv:1911.00068*.
- Pesce, E.; Withey, S. J.; Ypsilantis, P.-P.; Bakewell, R.; Goh, V.; and Montana, G. 2019. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Medical Image Analysis* 53: 26–38.
- Pham, H. H.; Le, T. T.; Tran, D. Q.; Ngo, D. T.; and Nguyen, H. Q. 2019. Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels. *arXiv preprint arXiv:1911.06475*.
- Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpankaya, K.; et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
- Shin, H.-C.; Roth, H. R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; and Summers, R. M. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging* 35(5): 1285–1298.
- Su, K.; and Geng, X. 2019. Soft facial landmark detection by label distribution learning. In *AAAI*, volume 33, 5008–5015.
- Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.
- Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2097–2106. IEEE.
- Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; and Summers, R. M. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, 9049–9058. IEEE.

- Wu, J.; Ruan, S.; Lian, C.; Mutic, S.; Anastasio, M. A.; and Li, H. 2018. Active learning with noise modeling for medical image annotation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 298–301. IEEE.
- Xu, N.; Lv, J.; and Geng, X. 2019. Partial label learning via label enhancement. In *AAAI*, volume 33, 5557–5564.
- Xu, N.; Tao, A.; and Geng, X. 2018. Label Enhancement for Label Distribution Learning. In *IJCAI*, 2926–2932.
- Yao, L.; Poblenz, E.; Dagunts, D.; Covington, B.; Bernard, D.; and Lyman, K. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*.
- Yao, L.; Prosky, J.; Poblenz, E.; Covington, B.; and Lyman, K. 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*.
- Ye, W.; Yao, J.; Xue, H.; and Li, Y. 2020. Weakly Supervised Lesion Localization With Probabilistic-CAM Pooling. *arXiv preprint arXiv:2005.14480*.
- Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; and Xu, D. 2020. When Radiology Report Generation Meets Knowledge Graph. In *AAAI*, 12910–12917.
- Zhou, B.; Li, Y.; and Wang, J. 2018. A weakly supervised adaptive densenet for classifying thoracic diseases and identifying abnormalities. *arXiv preprint arXiv:1807.01257*.
- Zhou, Y.; Zhou, T.; Zhou, T.; Fu, H.; Liu, J.; and Shao, L. 2021. Contrast-Attentive Thoracic Disease Recognition with Dual-Weighting Graph Reasoning. *IEEE Transactions on Medical Imaging* 1–1. doi:10.1109/TMI.2021.3049498.
- Zoph, B.; Vasudevan, V.; Shlens, J.; and Le, Q. V. 2018. Learning transferable architectures for scalable image recognition. In *CVPR*, 8697–8710. IEEE.