DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays

Bingzhi Chen [a], Jinxing Li [b,c], Xiaobao Guo [a], Guangming Lu [a,*]

[a] Harbin Institute of Technology, Shenzhen, China
[b] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Hong Kong
[c] Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

ABSTRACT

Recently, deep convolutional neural networks (DCNNs) such as the famous ResNet and DenseNet have achieved significant improvements in the field of automatic analysis of chest X-rays (CXRs). However, we observe that a wider network can combine characteristics from different DCNNs to improve the ability of object recognition compared with the single networks. In this paper, we focus on the cooperation and complementarity of dual asymmetric DCNNs and present a novel dual asymmetric feature learning network named DualCheXNet for multi-label thoracic disease classification in CXRs. Correspondingly, two asymmetric subnetworks based on the ResNet and DenseNet are combined to adaptively capture more discriminative features of different abnormalities from the raw CXRs. Specifically, the proposed method enables two different feature fusion operations, such as feature-level fusion (FLF) and decision-level fusion (DLF), which exactly form the complementary feature learning embedded in DualCheXNet. Moreover, an iterative training strategy is designed to integrate the loss contribution of the involved classifiers into a unified loss, and optimize the process of complementary features learning in an alternative way. Extensive experiments on the ChestX-ray14 dataset clearly substantiate the effectiveness of the proposed method as compared with the state-of-the-art baselines.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

In recent decades, medical imaging [1–4] technology has been widely applied in clinical medicine. As one of the most commonly used screening techniques in the diagnosis of many thoracic diseases (Atelectasis, Cardiomegaly, Pneumonia, etc.), chest X-ray imaging can provide significant on the clinical diagnosis and treatment. However, most chest X-ray images suffer from several unavoidable problems, such as the complicated background, multiple potential abnormalities and various interactions between abnormal patterns. Therefore, the clinical analysis of chest X-ray images (CXRs) continues to be a complicated process that relies on the manual annotation of expert radiologists. Moreover, the radiologists are required to expend considerable effort attempting to assess the quality of diagnosis results. To mitigate these problems and enhance the confidence of the radiologist, the automatic analysis of CXRs has become an increasingly important technology in clinical diagnosis.

Many traditional methods of automatic analysis techniques focus on the low-level features of abnormalities, such as appearance, texture, and contrast, etc., to make the predictions for the different pathologies. But not all spatial features in CXRs are available for the disease forecasting because some cluttered or "noisy" regions even generate distractions. Meanwhile, it is of great difficulty to detect some small abnormalities such as "Mass" and "Nodule" from their complicated background. In the past decade, DCNNs [5–8] have proven to be effective models for a variety of image classification tasks to explore high-level complex feature representations from the raw images. With the increased availability of large-scale annotated medical image archives, DCNNs have achieved the rapid and tremendous progress in a variety of automatic medical image analysis tasks, e.g. skin lesions detection [9,10], pulmonary tuberculosis detection [4,11], musculoskeletal abnormality classification [12] and brain tumor grade classification [13]. Similarly, recent works by Wang et al. [22]) and Rajpurkar et al. [24] have shown that the DCNNs actually outperform the traditional performance on thoracic disease classification in CXRs. Therefore, how to design a reasonable DCNNs model that can learn more effective features for detecting thoracic diseases has become the primary mission of this study.

* Corresponding author.
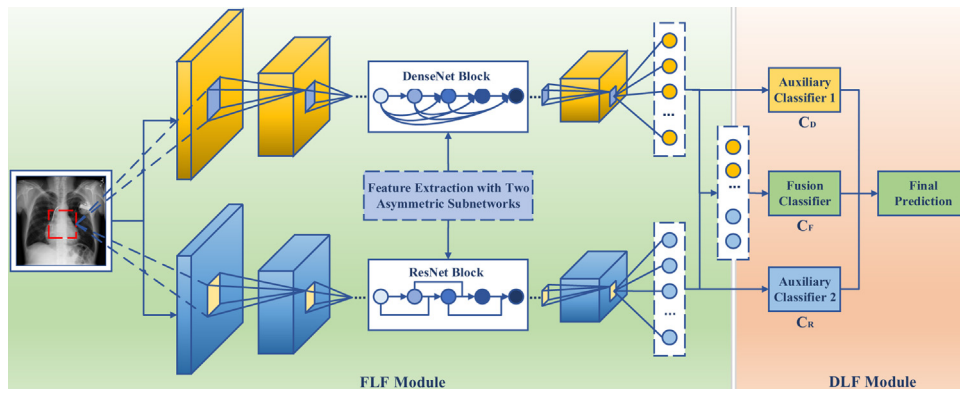E-mail address: luguangm@hit.edu.cn (G. Lu).

Fig. 1. Illustration of the proposed DualCheXNet. Two asymmetric subnetworks based on the DenseNet and ResNet are used to learn more complementary features from the same input image. The two-stream features are concatenated into the fusion classifier after extraction. And two auxiliary classifiers are appended behind the feature extractor of each subnetwork.

With the aim of extracting more high-level features, many state-of-the-art methods have made great efforts to construct a much deeper architecture. Driven by these works, DCNNs have evolved from the original AlexNet [14] and VGGNet [15] to the famous ResNet [16] and DenseNet [17]. Initially, AlexNet only consists of 8 layers, and VGGNet features 19 layers, but now ResNet and DenseNet have surpassed the 100-layer barrier. However, as the increasing of the model's depth, the resulting problems of vanishing gradient and overfitting have become prominent. Moreover, a single network with a deeper structure might drop some distinctive details in the middle layers, which are low-level but essential to the hard-classified abnormalities in CXRs. Such problems have become bottlenecks and impediments to the development of ImageNet [5] model's depth. Therefore, many researchers attempt to broaden the width of networks to enjoy the benefits of multiple DCNNs. For example, a novel framework named DualNet [18] is presented by Hou et al., that utilizes two isomorphic DCNNs to learn more accurate features complementary to each other. But the obtained features extracted from these two symmetric subnetworks might be duplicated. Moreover, many deep hashing methods [19,20] are proposed to perform simultaneous feature learning and hash-code learning with two parallel subnetworks. Inspired by these architectures, in this paper, we propose a novel dual asymmetric network named DualCheXNet to further boost the performance of multi-label thoracic disease classification. Here, our DualCheXNet consists of two asymmetric subnetworks, i.e. the ResNet and DenseNet, which is quite distinctive from the DualNet.

To tackle the vanishing gradient, both Resnet and DenseNet successfully adopt skip connections in their micro-blocks and create short paths from early layers to later layers, which enable the gradient to flow to the preceding layers well. With the inner structure design of the micro block, ResNet adds the input feature to output feature through its residual-blocks. Different from ResNet, in each dense-block of DenseNet, the input channels are concatenated with its outputs. Based on these properties, ResNet enables its residual-blocks to reuse the early features through the residual path, while DenseNet aims to urge its dense-blocks to explore new features through the densely connected path. Therefore, we can infer that the features extracted from ResNet and DenseNet are different and unique, which could be complemental in superiority for both of them. According to their property of advantage complementary, the proposed DualCheXNet is to integrate the advantages of ResNet and DenseNet into a wider architecture, to extract more discriminative features of different abnormalities from the raw CXRs. That is, an additional network is expected to learn more complementary details which are missing in the other. By contrast, a wider network can combine characteristics from different subnetworks to improve the ability of object recognition.

As shown in Fig. 1, we consider our DualCheXNet as an end-to-end framework, which is composed of two logical modules, i.e., the FLF module and DFL module. Specifically, the feature extractors of two subnetworks in the FLF module are located side by side, so as to form the unified feature extractor of DualCheXNet. Then these two asymmetric feature streams are concatenated into the following fusion classifier to perform a FLF optimization operation. Moreover, in the DLF module, two auxiliary classifiers are appended behind the end of their corresponding subnetworks to make the obtained features discriminative alone in the training phase. The main idea of the DLF module is to perform a DLF optimization operation by balancing the contribution of three involved classifiers, i.e., the fusion classifier and two auxiliary classifiers. In order to achieve this, an iterative training strategy is also designed to integrate the loss contribution of these classifiers into a unified loss of DualCheXNet, as described in Section 3. Thus, the complementarity between the learned features from two asymmetric subnetworks can be revealed through such unified loss.

The main contributions of the proposed method are as follows:

(1) This work is the first to focus on the cooperation and complementarity of dual asymmetric subnetworks used in the field of thoracic disease classification.
(2) A novel dual asymmetric feature learning network based on the DenseNet and ResNet is proposed to asymmetrically learn more complementary features from the raw CXRs for multi-label thoracic disease classification..
(3) An iterative training strategy based on the unified loss function is designed to fuse the decisions of three involved classifiers in an alternative way, which can better preserve the complementary features, and efficiently speeds up the convergence at the training phase.
(4) Experimental results on the ChestX-ray14 dataset show that the proposed DualCheXNet performs favorably against the previous state-of-the-art baselines.

The rest of the paper is organized as follows. Section 2 introduces the related work about thoracic disease classification in CXRs. The proposed method is described in Section 3. The comprehensive experiments are given in Section 4. And Section 5 concludes the whole work.

2. Related works

With the availability of large-scale annotated CXRs archives, many state-of-the-art methods have been proposed to develop the automatic analysis of CXRs. In this section, we briefly introduce the

related works of multi-label thoracic disease classification on the ChestX-ray14 dataset.

Initially, some researchers attempted to tackle this task by using classical ImageNet single models. For instance, Wang et al. [22] first announced the ChestX-ray14 dataset and investigated the ImageNet pre-trained DCNNs models, i.e., AlexNet, VGGNet, GoogLeNet [7], ResNet, to perform multi-label pathology classification on the ChestX-ray14 dataset. By fine-tuning a modified 121-layer DenseNet, Rajpurkar et al. [24] exploited their CheXNet to detect all 14 diseases in the ChestX-ray14 dataset, especially for "Pneumonia". But such single networks only have a simple response pattern. As DCNNs become increasingly deep, some original features from the raw input might still be ignored when they reach the end of the network. Therefore, a single network cannot really meet the requirements of multi-label thoracic disease classification on the ChestX-ray14 dataset. Next, some state-of-the-art works also focused on the relationships among image labels. For example, Yao et al. [23] utilized a variant of DenseNet and long short-term memory networks (LSTM) to learn the interdependencies among target labels and explore potential labels without pre-training.

Furthermore, some researchers introduced their attention mechanism to capture the discriminative features from the lesion regions. For example, Guan et al. [25] used attention guided DCNNs to improve the recognition performance by correcting image alignment and focusing the networks' attention on a reduced view. In practice, their network consists of two symmetric subnetworks with the different input images. Gundel et al. [26] proposed a location-aware Dense Networks to detect pathologies in CXRs, by fusing both the ChestX-ray14 and PLCO [27] datasets and using the spatial information in the high-resolution CXRs. Tang et al. [28] identified the thoracic disease category with an attention-guided curriculum learning method. Shen et al. [29] handled the thoracic disease identification and classification by combining the routing-by agreement mechanism and the DCNNs. However, such methods first require a significant amount of work to locate the position of lesion regions. More recently, Guan et al. [30] also proposed a category-wise residual attention learning (CRAL) framework for multi-label thorax disease classification, which can suppress the obstacles of irrelevant classes by endowing small weights to the corresponding feature representations. Despite the great improvement brought by these works, yet more workable techniques of complementary features learning remain unexplored. Different from those methods, the motivation of the proposed DualCheXNet in this paper aims to combine the advantages of two asymmetric subnetworks, i.e. the ResNet and DenseNet to adaptively captures more discriminative features for multi-label thoracic disease classification in CXRs.

As mentioned in [26], there is a great variability in the experimental results when we use different split standards of the ChestX-ray14 dataset, due to the class imbalance problem. Thus, it is unfair and problematic for some aforementioned works to show their best results by using their own non-official split standards. In order to illustrate the effectiveness of the proposed DualCheXNet, we make a comparison with other state-of-the-art baselines by grouping them according to their split standards, as shown in Section 4. Consequently, in this paper, we first follow strictly the official split standards of the ChestX-ray14 dataset, which was released by Wang et al. [22]. But for a more comprehensive comparison with other methods, we also evaluate the performance of DualCheXNet by using other publicly-available split standards.[1]

---

[1] Both the official split standards and non-official split standards of the ChestX-ray14 dataset used in this paper are published on GitHub. https://github.com/Binz-Chen/DualCheXNet.

3. The proposed method

In this section, we first give some notations used throughout this paper, as well as the problem definition. The details of the proposed DualCheXNet are then described, followed by its fusion operations and training strategy.

3.1. Notations and problem definition

Since there are two asymmetric streams in the proposed DualCheXNet, we first utilize the $I_D = \{d_1, \ldots, d_i, \ldots, d_N\} \in \mathbb{R}^{N \times l \times w \times 3}$ and $I_R = \{r_1, \ldots, r_i, \ldots, r_N\} \in \mathbb{R}^{N \times l \times w \times 3}$ to denote the input images of two subnetworks, respectively, where N is the number of training samples, l and w are the length and width of each input image. Note that I_D and I_R are the different representations of the same input images. The labels for each CXRs are expressed as a 14-dimensional label vector $x = \{x_1, \ldots, x_i, \ldots, x_N\}$, where $n = 14$ and $i \in \{0, 1, 2, \ldots, n\}$. $x_i \in \{0, 1\}$ represents the ground truth of the ith label, $x_i = 0$ indicates its absence while $x_i = 1$ indicates its presence. Such label representation transits the multi-label pathology classification task into a regression-like loss setting. The purpose of our two parallel DCNNs is to learn two different feature mapping outputs $f_D = \tau(I_D, x)$ and $f_R = \tau(I_R, x)$ to complement each other. Let the $C_F = \sigma(f_F) = \sigma(f_D \oplus f_R)$, $C_D = \sigma(f_D)$, $C_R = \sigma(f_R)$ denote the classification outputs of the fusion classifier and two auxiliary classifiers, where σ is a sigmoid function. Let the $L_F = \varphi(C_F)$, $L_D = \varphi(C_D)$, $L_R = \varphi(C_R)$ denote the loss of the fusion classifier and two auxiliary classifiers, where φ is the loss function used in each loss layer.

3.2. Dual asymmetric DCNNs architecture

The main framework of DualCheXNet is shown in Fig. 1, which mainly contains two modules, i.e. the FLF module and DLF module.

3.2.1. FLF module

In the FLF module, two asymmetric subnetworks based on the ResNet and DenseNet are placed side by side for, which exactly form the feature extractor of DualCheXNet. Then, two asymmetric feature streams are concatenated into the following fusion classifier. Noted that the obtained features after FLF optimization are supposed to be the more discriminative and informative compared to the features extracted by each subnetwork. In this paper, we investigate two different configurations of the FLF module which are denoted as DualCheXNet-1 and DualCheXNet-2. More specifically, these two involved subnetworks are initialized with the corresponding DenseNet and ResNet pre-trained models, i.e., DualCheXNet-1 is initialized with the 121-layer DenseNet and 50-layer ResNet, while DualCheXNet-2 is initialized with the 169-layer DenseNet and 101-layer ResNet.

3.2.2. DLF module

Three classifiers such as the fusion classifier C_F and two auxiliary classifiers C_D and C_R are involved in the DLF module, which aim to make the features provided from each subnetwork discriminative alone. Fig. 2 illustrates the structure of the classifier in the DLF module. Moreover, in each classifier, a sigmoid layer is directly appended behind the final FC layer to normalize feature mapping outputs f ($f \in \{f_D, f_R, f_F\}$), and obtain the classification scores,

$$p(i|I) = 1/(1 + \exp(-p(i|f))) \tag{1}$$

where $p(i|I)$ indicates the probability score of the ith label.

During the training phase, we noticed that the ChestX-ray14 is not only subjected to the class imbalance problem of the positives and negatives, but also suffers a huge imbalance between
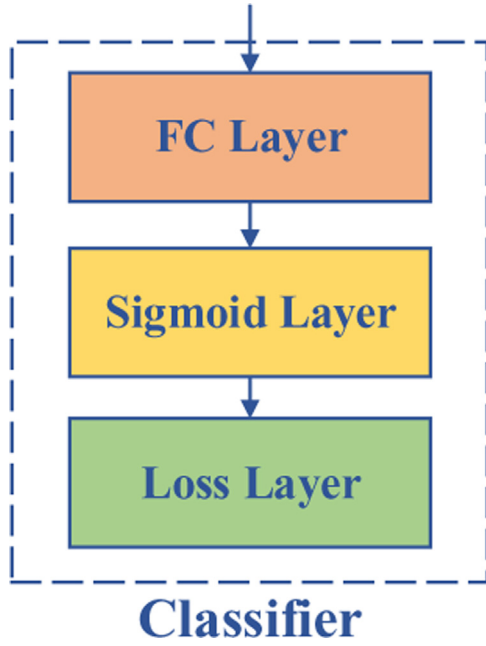
Fig. 2. Illustration of the classifiers in the decision-level module. A sigmoid layer and the proposed loss layer are appended to the final fully connected (FC) layer in sequence.

the easily-classified samples and hard-classified samples. as mentioned in [22], some of them such as "Mass" and "Pneumonia" are difficult to identify because of their huge within-class appearance variations. Lin et al. [21] have evidenced that vast easily-classified negatives can comprise the majority of loss and dominate the gradient, both of which result in performance degradation. To mitigate these problems, the feature extractors are supposed to focus more on the hard-classified samples, and reducing the overload of vast easily-classified negatives. Therefore, we introduced their proposed modulating factor $(1 - p(i|I))^{\lambda}$ into the weighted cross entropy (W-CE) loss function used in each loss layer. The variant of the W-CE loss function is defined as follows,

$$L(p(i|I), \boldsymbol{x}) =$$

$$-\frac{1}{|P|}\sum_{x_i=1}(1-p(i|I))^{\lambda}\ln(p(i|I)) - \frac{1}{|N|}\sum_{x_i=0}p(i|I)^{\lambda}\ln(1-p(i|I)) \quad (2)$$

where $|P|$ and $|N|$ are the total number of positive labels and negative labels in a batch of training set, and $(1 - p(i|I))^{\lambda}(\lambda \in [0, 2])$ is the selected modulating factor to balance the influence between the hard-classified and easily-classified samples in the training phase. When an existing abnormality is difficult to distinguish, and its corresponding probability score $(p(i|I))$ close 0. Correspondingly, the $(1 - p(i|I))^{\lambda}$ is near 1, so that its loss contribution would be strengthened, and vice versa.

3.3. Fusion optimization

Correspondingly, two different types of fusion operations i.e., the FLF optimization and DLF optimization are mainly discussed.

3.3.1. FLF optimization
In order to learn more complementary features, two asymmetric subnetworks f_R and f_D in the FLF module are first to models the difference among categories and discriminatively represent the raw inputs. And the features extracted from these two streams are then sent into the following fusion classifier for the complementary feature learning. Only the Concat fusion method is applied in

our FLF optimization operation. In this way, two asymmetric feature streams are directly concatenated to form more discriminative fused features.

3.3.2. DLF optimization
As mentioned above, all classifiers in our DLF module can make the learned features discriminative alone, and stimulate one another for common development. In order to further boost the performance of joint finetuning of the three involved classifiers, the key of our DLF optimization operation is to assemble the contribution of three involved classifiers with several simple fusion methods, such as AVG, MAX, and Concat, as shown in Eq. (3).

$$L = \begin{cases} \boldsymbol{avg}(L_F, L_D, L_r), & AVG \\ \boldsymbol{max}(L_F, L_D, L_r), & MAX \\ \boldsymbol{concat}(L_F, L_D, L_r), & Concat \end{cases} \quad (3)$$

In our experiments, we found that when we concatenate the involved classifiers with their corresponding weights, the proposed DualCheXNet can achieve the best performance. Thus, the proposed unified loss of DualCheXNet is defined as:

$$L = w_F L_F + w_D L_F + w_R L_F \quad (4)$$

where w_F, w_D, w_R are the loss weights of the three classifiers $\boldsymbol{C_F}, \boldsymbol{C_D}$ and $\boldsymbol{C_R}$. In our experiments, w_F, w_D and w_R are designed to balance the processes of joint finetuning, where w_F are set to 1 while w_D and w_R are set to 0.3 for the best performance.

Algorithm 1. Dual asymmetric features learning method

Input data: Training data $(\boldsymbol{I}, \boldsymbol{x})$
Output: The outputs of the fusion classifier and two auxiliary classifiers $\boldsymbol{C_F}, \boldsymbol{C_D}$ and $\boldsymbol{C_R}$
Initialization: Initialize weights of the DualCheXNet by using the DenseNet and ResNet pretrained ImageNet model; the last FC layer in each classifier is initialized randomly.
1: **while** not converged or the maximum iteration is not reached **do**
2: **Update the DenseNet Stream:** Fixing the ResNet stream, and optimize the DenseNet stream, fusion classifier $\boldsymbol{C_F}$, and auxiliary classifiers $\boldsymbol{C_D}$ according to the loss function Eq. (5).
3: **Update the ResNet Stream:** Fixing the DenseNet stream, and optimize the ResNet stream, fusion classifier $\boldsymbol{C_F}$, and auxiliary classifiers $\boldsymbol{C_R}$ according to the loss function Eq. (6).
4: **Update the whole network:** Optimize the DenseNet and ResNet streams, fusion classifier $\boldsymbol{C_F}$ and two auxiliary classifiers $\boldsymbol{C_D}$ and $\boldsymbol{C_R}$ according to the loss function Eq. (4).
5: **EndWhile**

3.4. Iterative training strategy

In order to reduce GPU memory usage and overcome the problem of overfitting excellently, we design an efficient iterative training strategy to fuse the decisions of the three classifiers in an alternative way and optimize the two asymmetric streams alternatively. Moreover, in each iteration, one subnetwork can extract more complementary features of abnormalities to yield more discriminative fused features, when the other remains unchanged. Overall, the optimization of the proposed iterative training strategy is presented in Algorithm 1. It can effectively improve the generalization ability of the proposed DualCheXNet. The training process is as follows:

3.4.1. Update the DenseNet stream with the ResNet stream fixed
By fixing the ResNet stream, the loss function Eq. (4) can be transformed to

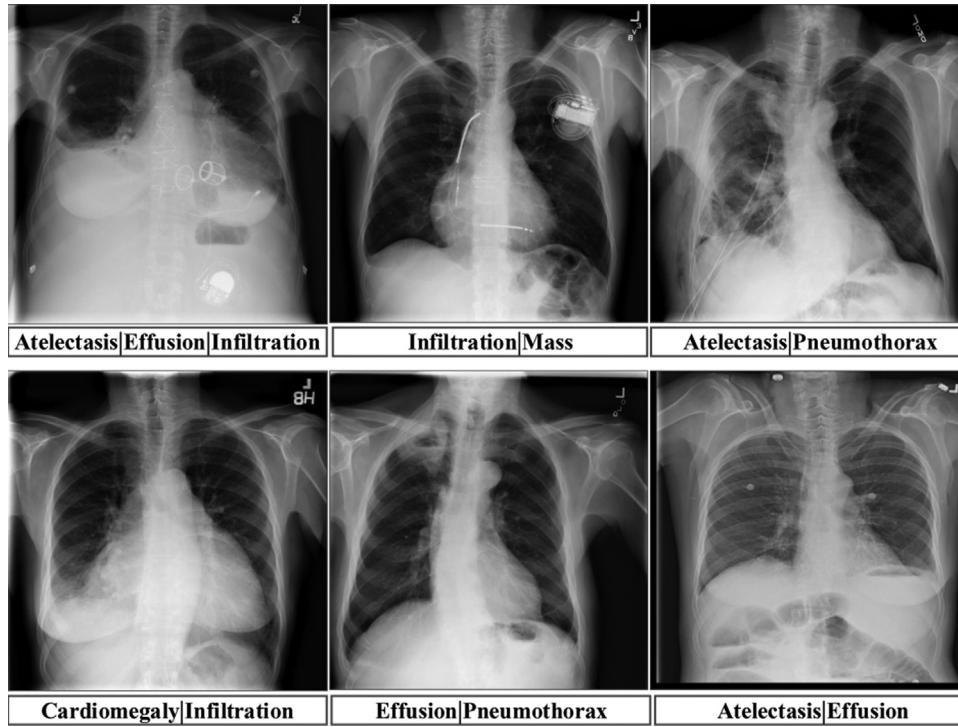$$L = w_F \boldsymbol{L_F} + w_D \boldsymbol{L_D} \quad (5)$$

Fig. 3. Six multi-label samples from the ChestX-ray14 dataset. It was noted above that most positives CXRs in the ChestX-ray14 dataset are interpreted to indicate multiple present or suspected pathologies.

The DenseNet stream, fusion classifier C_F and auxiliary classifier C_D are updated.

3.4.2. Update the ResNet stream with the DenseNet stream fixed

By fixing the DenseNet stream, the loss function Eq. (4) can be transformed to:

$$L = w_F L_F + w_R L_R \qquad (6)$$

The ResNet stream, fusion classifier C_F and auxiliary classifier C_R are updated.

3.4.3. Update both the DenseNet and ResNet streams

The whole DualCheXNet is updated according to the loss function Eq. (4).

4. Experiments

In this section, the proposed DualCheXNet is evaluated on ChestX-ray14 dataset to detect all 14 diseases. We first give a description of the dataset, followed by some related baselines. After that is the evaluation metric and implementation settings used in this paper. Then a testifying check of the analyses of each parameter's influence further enhances the reliability of our method. At last, comparisons with the state-of-the-art methods are presented based on the official and non-official split standards, respectively.

4.1. Dataset

The ChestX-ray14 dataset is a large-scale chest X-ray images database with 14 common thoracic disease categories. It contains 112,120 frontal-view X-ray images from 32,717 patients. Fig. 3 illustrates six samples in the ChestX-ray14 dataset with multiple pathologies. As shown in Table 1, a wide range of abnormalities are distributed in the ChestX-ray14 dataset, such as "Atelectasis", "Nodule" "Pneumonia", etc., but they are highly unequal. For

Table 1
The total number of images across all abnormalities in the ChestX-ray14 dataset.

Items	Labels	Items	Labels
Atelectasis	11559	Consolidation	4667
Cardiomegaly	2776	Edema	2303
Effusion	13317	Emphysema	2516
Infiltration	19894	Fibrosis	1686
Mass	5782	P_T	3385
Nodule	6331	Hernia	227
Pneumonia	1431	**No finding**	60412
Pneumothorax	5302		

example, 60,412 samples are normal, which are labeled with "No Finding", while only 227 samples are labeled with "Hernia".

4.2. Baseline

We divide compared approaches into two categories, based on whether they use the official split standards of the ChestX-ray14 dataset. To demonstrate the superiority of the proposed DualCheXNet, some existing methods in Group 1 are based on the official split standards and used for comparison, including U-DCNN [22], DNet [26], AGCL [28], DR_DNN [29] and, CRAL [30]. Moreover, we also compare with some state-of-the-art methods in Group 2, i.e., LSTM-DNet [23], DNet [26] and CheXNet [24], which are conducted with other non-official split standards. Since the test sets of LSTM-DNet and CheXNet are not selected strictly on the basis of the official split standards of the ChestX-ray14 dataset, the comparison of this group is for reference only.

4.3. Evaluation protocol and implementation

4.3.1. Evaluation protocol

In our experiment, both the official and non-official split standards randomly shuffle the dataset into three subsets: 70% for training, 10% for validation and 20% for testing. Consistent with
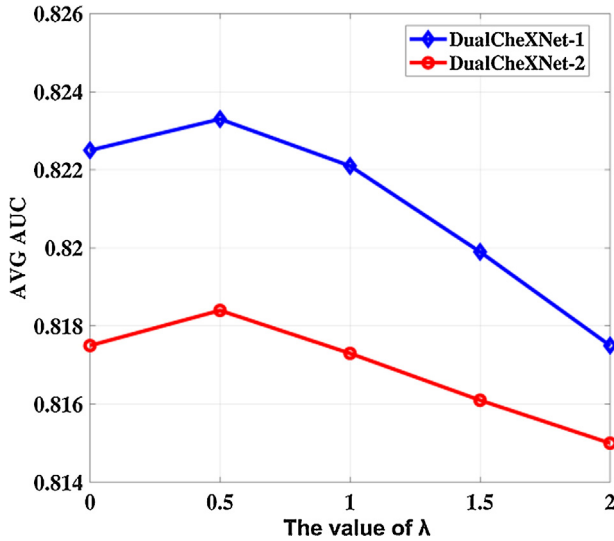
Fig. 4. Average AUC scores of the DualCheXNet-1 and DualCheXNet-2 with different settings of λ.

previous works, the area under the roc curves (AUC) [31] is voted as the only evaluation metric to evaluate and validate the performance of the proposed DualCheXNet for all 14 thoracic diseases classification.

4.3.2. Implementation

We implement the proposed DualCheXNet with the deep learning toolbox PyTorch on 4 TITAN XP GPUs. During the training, we optimize the network by using stochastic gradient descent (SGD) [32] with a mini-batch size of 8, and the learning rate is $lr = 0.001$, which will be reduced by a factor of 10 when the validation loss reaches a plateau. In general, a higher resolution image input is necessary to faithfully represent those small and localized abnormalities. For the data augmentation in our experiments, we first resize the original images from 1024×1024 pixels to 556×556 pixels, and randomly crop it to 512×512 pixels.

4.4. Parameter analysis

4.4.1. Hyperparameters λ

The λ is a key hyperparameter of the proposed unified loss function that controls the strength of the modulating factor $(1 - p(i|\mathbf{I}))^{\lambda}$ in Eq. (2). By fixing other hyperparameters, a range of values of λ ($\lambda \in \{0, 0.5, 1.0, 1.5, 2.0\}$) are tested, as shown in Fig. 4. When $\lambda = 0$, our loss is equal to a normal W-CE loss. As λ increases, on the one hand, the loss contribution of hard-classified samples is further enhanced that urges the feature extractors to focus more on these samples. On the other hand, the loss contribution of easily-classified negatives gets further discounted that reduces the overload of vast easily-classified negatives in the training phase. Compared with the W-CE loss, the proposed loss can boost the performance of DualCheXNet. Compared with the W-CE loss, the unified loss is more suitable for the proposed method. In particular, both DualCheXNet-1 and DualCheXNet-2 achieve the best performance when λ is 0.5.

4.4.2. Hyperparameters w_F, w_D, w_R

Furthermore, the w_F, w_D and w_R introduced in Eq. (4) are applied to investigate the correlation of fusion classifier C_F and two auxiliary classifiers C_D and C_R in the training phase. By fixing the values of λ ($\lambda = 0.5$), different fusion methods are conducted to determine the best parameter combinations of w_F, w_D and w_R. As illustrated in Fig. 5, it turns out that the Concat fusion meth-
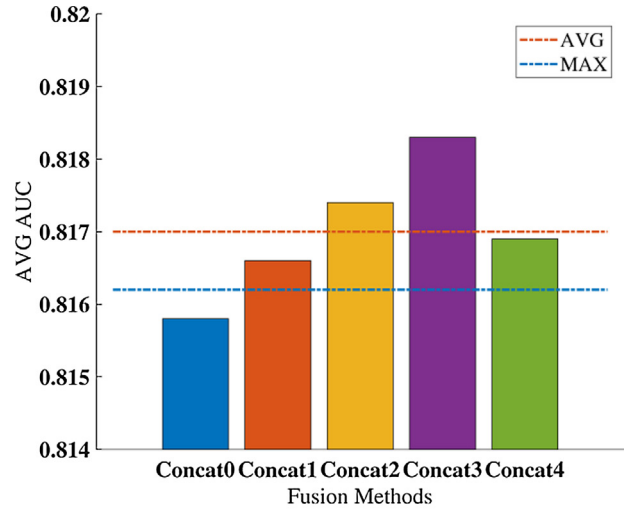
(a) DualCheXNet-1
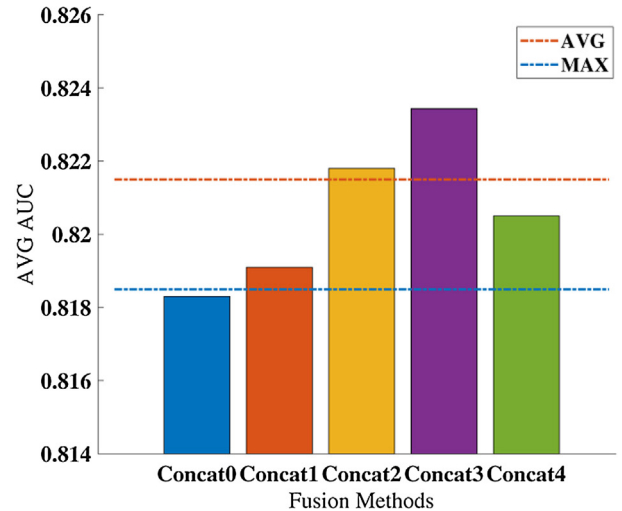
(b) DualCheXNet-2

Fig. 5. Average AUC scores of DualCheXNet-1 (a) and DualCheXNet-2 (b) with different fusion methods, including AVG, MAX, Concat0 ($w_F = 1.0$, $w_D = w_R = 0$), Concat1 ($w_F = 1.0$, $w_D = w_R = 0.1$), Concat2 ($w_F = 1.0$, $w_D = w_R = 0.2$), Concat3 ($w_F = 1.0$, $w_D = w_R = 0.3$), and Concat4 ($w_F = 1.0$, $w_D = w_R = 0.4$).

ods, such as Concat2 and Concat3, are more practical than AVG and MAX to learn more complementary features. And the setting Concat3 ($w_F = 1$, $w_D = w_R = 0.30$) can contribute to the best performance of DualCheXNet-1 and DualCheXNet-2 (the average AUC scores of DualCheXNet-1 and DualCheXNet-2 are 0.823 and 0.818, respectively.).

4.5. Evaluation

In this subsection, we take further experiments on the ChestX-ray14 dataset to analyze the performance of our method. The corresponding AUC and ROC curves [33] are presented. We first showcase the performance of the proposed DualCheXNet by using the official split standards and non-official split standards, respectively. By comparing with the corresponding single networks, the feasibility of the proposed DualCheXNet is demonstrated preliminarily. Subsequently, we also compare DualCheXNet with the state-of-the-art baselines according to their split standards. At last, the time cost of the proposed method when a testing sample is
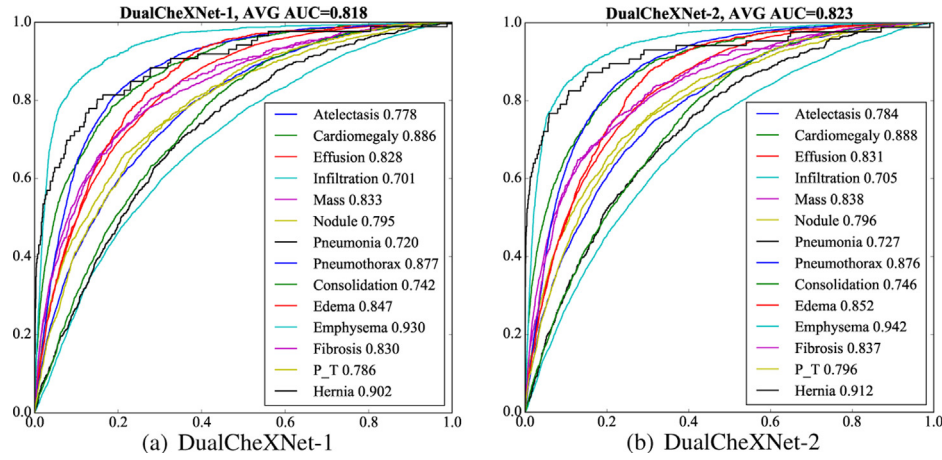
Fig. 6. ROC curves of DualCheXNet-1 (a) and DualCheXNet-2 (b) on all 14 pathologies using the official split standards. The AUC scores of each pathology are also presented in the figure.
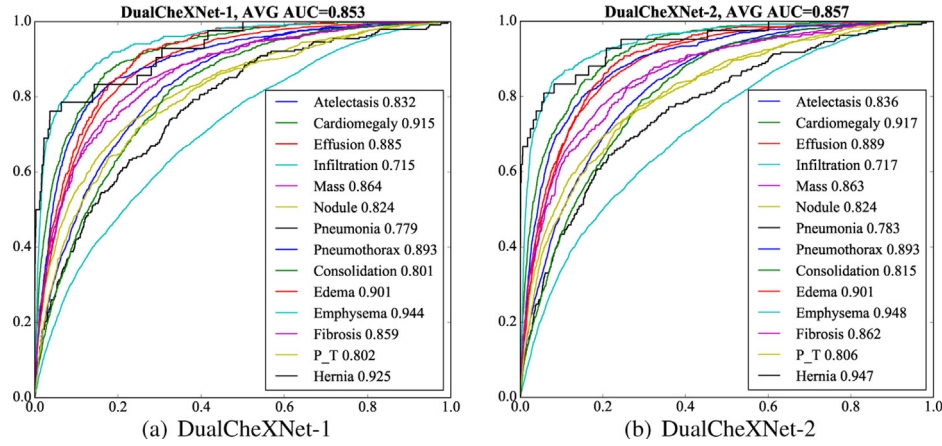

Fig. 7. ROC curves of DualCheXNet-1 (a) and DualCheXNet-2 (b) on all 14 pathologies using the non-official split standards. The AUC scores of each pathology are also presented in the figure.

given. Note that the best results are highlighted in bold for each row.

4.5.1. Comparison between different DualCheXNet structures

We first report the performance of two basic DualCheXNet structures, i.e., DualCheXNet-1 and DualCheXNet-2. Results are summarized in Fig. 6 and Fig. 7. Obviously, there is a huge gap between the performance of DualCheXNet using the official split standards and non-official split standards of the ChestX-ray14 dataset, respectively. For example, the average AUC scores of DualCheXNet-1and DualCheXNet-2 using the non-official split standards are 0.853 and 0.857, however, it is greatly reduced to 0.818 and 0.823 when we follow strictly the official split standards. This phenomenon indicates that the split standards of the dataset have a direct effect on classification performance. Overall, the experimental results show that DualCheXNet-2 outperforms DualCheXNet-1. Based on the official split standards, DualCheXNet-2 yields an average AUC score of 0.823, which is superior to the performance of DualCheXNet-1 (0.818). Especially, the AUC scores of "Emphysema", "P_T" and "Hernia" with DualCheXNet-2 are higher than DualCheXNet-1 by over 0.01, while the AUC scores of "Nodule", "Pneumonia", and "Fibrosis" with DualCheXNet-1 exceeds DualCheXNet-1 over 0.005. And the AUC scores of other pathologies are very close to or even higher than DualCheXNet-1. Besides, a low number of labels, such as "Hernia", "Emphysema" and "Cardiomegaly", yields high AUC scores that over 0.880. It is

probably because these pathologies usually have a specific shape of appearance in lesion areas, which are advantageous for identification. By contrast, DualCheXNet performs worse on the recognition of "Infiltration" than on tests in any other pathologies (0.701 and 0.705 for DualCheXNet-1 and DualCheXNet-2). We suspect this is because the various textures distributed in the lung area are difficult to identify.

4.5.2. Comparison with single networks

To prove the effectiveness of the proposed dual asymmetric deep structure, we compare the proposed DualCheXNet with their corresponding single networks. As shown in Table 2, experimental results demonstrate that the performance of the proposed DualCheXNet is consistently better than DenseNet and ResNet. Specifically, DualCheXNet-1 yields a higher average AUC score over the corresponding ResNet-50 (0.810) and DenseNet-121 (0.813), while DualCheXNet-2 yields a higher average AUC score over the corresponding ResNet-101 (0.812) and DenseNet-169 (0.815). Compared with the single networks, DualCheXNet-1 and DualCheXNet-2 achieve better performance in 7 and 11 pathologies, respectively, especially for the recognition of "Nodule", "Pneumothorax" and "P_T". The AUC scores of these three pathologies with DualCheXNet-2 are higher than their single networks by over 0.01. As further evidence, we also found that the performance of the ResNet and DenseNet have their own advantages and certain mutual supplement, which can verify the motivation of comple-

Table 2
Comparison of the AUC scores between DualCheXNet and the corresponding single networks on the official ChestX-ray14 test set. Both DualCheXNet-1 and DualCheXNet-2 outperform their corresponding single networks. 'D' and 'R' represent the DenseNet and ResNet, respectively. For each row, the best results are highlighted in bold.

CNN	R-50	D-121	DualCheXNet-1	R-101	D-169	DualCheXNet-2
Atelectasis	0.776	0.772	**0.778**	0.775	0.779	**0.784**
Cardiomegaly	**0.889**	0.881	0.886	**0.893**	0.882	0.888
Effusion	0.828	0.824	**0.828**	0.829	0.822	**0.831**
Infiltration	0.695	0.691	**0.701**	0.697	0.701	**0.705**
Mass	0.822	**0.835**	0.833	0.824	0.837	**0.838**
Nodule	0.764	0.772	**0.795**	0.766	0.775	**0.796**
Pneumonia	0.713	**0.721**	0.720	0.710	0.718	**0.727**
Pneumothorax	0.861	0.872	**0.877**	0.858	0.864	**0.876**
Consolidation	**0.745**	0.738	0.742	**0.749**	0.740	0.746
Edema	0.847	0.840	**0.847**	0.848	0.840	**0.852**
Emphysema	0.904	**0.938**	0.930	0.910	0.933	**0.942**
Fibrosis	0.821	0.822	**0.830**	0.818	0.824	**0.837**
P_T	0.762	0.774	**0.786**	0.772	0.781	**0.796**
Hernia	**0.914**	0.897	0.902	**0.924**	0.907	0.912
AVG AUC	0.810	0.813	**0.818**	0.812	0.815	**0.823**

Table 3
Comparison of the AUC scores between DualCheXNet and other published baselines on the official ChestX-ray14 test set. '*' represents the network used in the corresponding reference is not illustrated. For each row, the best results are highlighted in bold.

Group 1 CNN	U-DCNN [22] R-50	DR_DNN [29] *	AGCL [28] *	DNet [26] D-121	CRAL [30] D-121 &D-121	DualCheXNet-1 R-50 &D-121	DualCheXNet-2 R-101 &D-169
Atelectasis	0.700	0.766	0.756	0.767	0.781	0.778	**0.784**
Cardiomegaly	0.810	0.801	0.887	0.883	0.880	0.886	**0.888**
Effusion	0.759	0.797	0.819	0.828	0.829	0.828	**0.831**
Infiltration	0.661	0.751	0.689	**0.709**	0.702	0.701	0.705
Mass	0.693	0.760	0.814	0.821	0.834	0.833	**0.838**
Nodule	0.669	0.741	0.755	0.758	0.773	0.795	**0.796**
Pneumonia	0.658	**0.778**	0.729	0.731	0.729	0.720	0.727
Pneumothorax	0.799	0.800	0.850	0.846	0.857	**0.877**	0.876
Consolidation	0.703	**0.787**	0.728	0.745	0.754	0.742	0.746
Edema	0.805	0.820	0.848	0.835	0.850	0.847	**0.852**
Emphysema	0.833	0.773	0.906	0.895	0.908	0.930	**0.942**
Fibrosis	0.786	0.765	0.818	0.818	0.830	0.830	**0.837**
P_T	0.684	0.759	0.765	0.761	0.778	0.786	**0.796**
Hernia	0.872	0.748	0.875	0.896	**0.917**	0.902	0.912
AVG AUC	0.745	0.775	0.803	0.807	0.816	0.818	**0.823**

mentary feature learning used in our DualCheXNet. For example, DenseNet-121 outperforms ResNet-50 for the recognition of "Consolidation" (0.745 vs. 0.738), but they encounter the opposite for the recognition of "Emphysema" (0.904 vs. 0.938). Based on their property of advantage complementary, the proposed method can learn more discriminative features for these pathologies through complementary features learning (0.742 and 0.930 for the recognition of "Consolidation" and "Emphysema" with DualCheXNet-1). From the view of the model' depth, even though DenseNet-169 has a deeper structure, it also performs worse than our DualCheXNet-1 (0.815 vs.0.818), which further proves the effectiveness of the proposed method.

4.5.3. Comparison with state-of-the-art baselines

We also compare the proposed method with previous state-of-the-art baselines on the Chest X-ray14 dataset. The results are summarized in Tables 3 and 4 . For a fair comparison, we first compare the methods in Group 1 which utilize the public official split standards. As shown in Table 3, the proposed method reaches a new state of the art: the average AUC score for 14 pathologies is 0.823. Specifically, both DualCheXNet-1 (0.818) and DualCheXNet-2 (0.823) achieve superior performance over the previous state-of-the-art baselines. In particular, DualCheXNet-2 achieves an improvement of 7.8% compared with the U-DCNN (0.745) [22]. Moreover, DualCheXNet also outperforms CRAL (0.816) [30], which is the latest state-of-the-art baseline provided by Guan et al. And the AUC scores of for 12 pathologies with DualCheXNet are higher than CRAL, especially for the "Nodule"

(0.773 vs. 0.796), "Pneumothorax" (0.857 vs. 0.876), "Emphysema" (0.908 vs. 0.942), and "P_T" (0.778 vs. 0.796). Table 4 illustrates some baselines in Group 2 based on the non-official split standards. Note that the validity of the comparison in this group remains open to question. Both DualCheXNet-1 (0.853) and DualCheXNet-2 (0.857) outperform the these baselines, especially LSTM-DNet (0.798) [23] and U-DCNN (0.738) [22] with improvements of 5.9% and 11.9%. Moreover, the proposed method achieves better performance on 11 pathologies compared with CheXNet [24]. Overall, the classification performance of the proposed method is more practical than the previous baselines.

4.5.4. Test-time detection

Table 5 illustrates the comparison of the average test cost between DualCheXNet and previous baselines. The test is conducted on one TITAN XP GPU, where the time cost is recorded for 25,596 test images from the ChestX-ray14 test set and averaged at the end of the test run. Since the proposed method is structured with two asymmetric deep subnetworks, the time cost of DualCheXNet-1 (0.0432) and DualCheXNet-2 (0.0453) are slightly increased, but it can be negligible in clinical diagnosis. By contrast, the proposed method is worthwhile for the performance improvement of thoracic disease classification on the ChestX-ray14 dataset.

4.6. Qualitative results

Twelve classification results are presented in Fig. 8. The top-8 probability scores are presented for each sample, and the

Table 4
Comparison of the AUC scores between DualCheXNet and other published baselines on the non-official ChestX-ray14 test set. '#' represents the combination of the LSTM and DenseNet used in [23]. For each row, the best results are highlighted in bold.

Group 2 CNN	U-DCNN [22] R-50	LSTM-DNet [23] #	DNet [26] D-121	CheXNet [24] D-121	DualCheXNet-1 R-50 &D-121	DualCheXNet-2 R-101 &D-169
Atelectasis	0.716	0.772	0.826	0.809	0.832	**0.836**
Cardiomegaly	0.807	0.904	0.911	**0.925**	0.915	0.917
Effusion	0.784	0.859	0.885	0.864	0.885	**0.889**
Infiltration	0.609	0.695	0.716	**0.735**	0.715	0.717
Mass	0.706	0.792	0.854	**0.868**	0.864	0.863
Nodule	0.671	0.717	0.774	0.780	0.824	**0.824**
Pneumonia	0.633	0.713	0.765	0.768	0.779	**0.783**
Pneumothorax	0.806	0.841	0.872	0.889	0.893	**0.893**
Consolidation	0.708	0.738	0.806	0.790	0.801	**0.815**
Edema	0.835	0.882	0.892	0.888	0.901	**0.901**
Emphysema	0.815	0.829	0.925	0.937	0.944	**0.948**
Fibrosis	0.769	0.767	0.820	0.805	0.859	**0.862**
P_T	0.708	0.765	0.785	**0.806**	0.802	0.806
Hernia	0.767	0.914	0.941	0.916	0.925	**0.947**
AVG AUC	0.738	0.798	0.841	0.841	0.853	**0.857**

Table 5
Comparison of the average test cost between DualCheXNet and previous baselines on ChestX-ray14 test set.

Baselines CNN	U-DCNN [22] R-50	DNet [26] D-121	CheXNet [24] D-121	ResNet-110 [16] R-101	DenseNet-169 [17] D-169	DualCheXNet-1 R-50 & D-121	DualCheXNet-2 R-101 & D-169
Time (s)	0.0304	0.0352	0.0359	0.0317	0.0381	0.0432	0.0453

Fig. 8. Examples of classification results. The top-8 predicted categories and the corresponding probability scores are presented. The ground-truth pathologies are highlighted in red.

ground-truth pathologies are highlighted in red. For the multi-label pathology classification task in CXRs, we can see that the probability scores of ground truth pathologies are on the top. Moreover, there are large gaps between the scores of ground truth pathologies and other pathologies, e.g., in row 1, column 3, the prediction of ground truth pathology ("Infiltration" is 0.717) achieves the highest score, which far exceeds other unrelated pathologies ("Nodule" is 0.171, "Effusion" is 0.116, and "Atelectasis" is 0.110).

5. Conclusion

In this paper, we propose a novel dual asymmetric feature learning network named DualCheXNet for multi-label thoracic disease classification in CXRs. Specifically, two asymmetric subnetworks based on the DenseNet and ResNet are coordinated to integrate the complementary features learning into the proposed DualCheXNet, in order to extract more discriminative features of different abnor-

malities from the raw CXRs. Compared with other baselines, the proposed method reaches a new state of the art: the average AUC score for 14 pathologies is 0.823. Therefore, the proposed method is worthwhile for the performance improvement of thoracic disease classification on the ChestX-ray14 dataset. Our research shows that the complementary feature learning with two asymmetric DCNNs is also beneficial to multi-label medical image classification, yet more sophisticated techniques remain unexplored, it is worth discussing in the future.

Acknowledgments

References

[1] W. Ayadi, W. Elhamzi, I. Charfi, M. Atri, A hybrid feature extraction approach for brain MRI classification based on bag-of-words, Biomed. Signal Process. Control 48 (2019) 144–152.
[2] D. Gupta, R.S. Anand, B. Tyagi, A hybrid segmentation method based on Gaussian kernel fuzzy clustering and region based active contour model for ultrasound medical images, Biomed. Signal Process. Control 16 (2015) 98–112.
[3] M.J. Willemink, M. Persson, A. Pourmorteza, N.J. Pelc, D. Fleischmann, Photon-counting CT: technical principles and clinical prospects, Radiology 289 (2) (2018) 293–312.
[4] P. Lakhani, B. Sundaram, Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks, Radiology 284 (2017) 574–582.
[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks., Advances in Neural Information Processing Systems (2012) 1097–1105.
[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2015, pp. 1–9.
[8] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 1440–1448.
[9] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.
[10] S. Pathan, K.G. Prabhu, P.C. Siddalingaswamy, Techniques and algorithms for computer aided diagnosis of pigmented skin lesions – a review, Biomed. Signal Process. Control 39 (2018) 237–262.
[11] S. Stirenko, Y. Kochura, O. Alienin, O. Rokovyi, Y. Gordienko, P. Gang, W. Zeng, Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation, 2018 IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO (2018) 422–428.
[12] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, C. Langlotz, Mura dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs, 2017, arXiv preprint arXiv:1712.06957.
[13] G. Mohan, M.M. Subashini, MRI based medical image analysis: survey on brain tumor grade classification, Biomed. Signal Process. Control 39 (2018) 139–161.
[14] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 770–778.
[17] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2261–2269.
[18] S. Hou, X. Liu, Z. Wang, DualNet: learn complementary features for image recognition, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 502–510.
[19] Q. Li, Z. Sun, R. He, T. Tan, Deep supervised discrete hashing, Advances in Neural Information Processing Systems (2017) 2482–2491.
[20] W.J. Li, S. Wang, W.C. Kang, Feature learning based deep supervised hashing with pairwise labels, in: IJCAI International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2016, pp. 1711–1717.
[21] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2999–3007.
[22] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 3462–3471.
[23] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, 2017, arXiv preprint arXiv:1710.10501.
[24] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, M.P. Lungren, Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
[25] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification, 2018, arXiv preprint arXiv:1801.09927.
[26] S. Guendel, S. Grbic, B. Georgescu, K. Zhou, L. Ritschl, A. Meier, D. Comaniciu, Learning to recognize abnormalities in chest X-rays with location-aware dense networks, 2018, arXiv preprint arXiv:1803.04565.
[27] J.K. Gohagan, P.C. Prorok, R.B. Hayes, B.S. Kramer, The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National Cancer Institute: history, organization, and status, Control. Clin. Trials 21 (6) (2000) 251S–272S.
[28] Y. Tang, X. Wang, A.P. Harrison, L. Lu, J. Xiao, R.M. Summers, Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 249–258.
[29] Y. Shen, M. Gao, Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 389–397.
[30] Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recognit. Lett. (2018).
[31] J.M. Lobo, A. Jiménez-valverde, R. Real, AUC: a misleading measure of the performance of predictive distribution models, Global Ecol. Biogeogr. 17 (2008) 145–151.
[32] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT 2010-19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers, Springer Berlin, 2010, pp. 177–186.
[33] L.Y.H. Liu, L.Y.H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.