Cognitive Computation (2022) 14:1362–1377
https://doi.org/10.1007/s12559-022-10032-4

MXT: A New Variant of Pyramid Vision Transformer for Multi-label Chest X-ray Image Classification

Xiaoben Jiang[1] · Yu Zhu[1] · Gan Cai[1] · Bingbing Zheng[1] · Dawei Yang[2,3]

Received: 26 October 2021 / Accepted: 26 May 2022 / Published online: 3 June 2022

Abstract

Nowadays, the global COVID-19 situation is still serious, and the new mutant virus Delta has already spread all over the world. The chest X-ray is one of the most common radiological examinations for screening catheters and diagnosis of many lung diseases, which plays an important role in assisting clinical diagnosis during the outbreak. This study considers the problem of multi-label catheters and thorax disease classification on chest X-ray images based on computer vision. Therefore, we propose a new variant of pyramid vision Transformer for multi-label chest X-ray image classification, named MXT, which can capture both short and long-range visual information through self-attention. Especially, downsampling spatial reduction attention can reduce the resource consumption of using Transformer. Meanwhile, multi-layer overlap patch (MLOP) embedding is used to tokenize images and dynamic position feed forward with zero paddings can encode position instead of adding a positional mask. Furthermore, class token Transformer block and multi-label attention (MLA) are utilized to offer more effective processing of multi-label classification. We evaluate our MXT on Chest X-ray14 dataset which has 14 disease pathologies and Catheter dataset containing 11 types of catheter placement. Each image is labeled one or more categories. Compared with some state-of-the-art baselines, our MXT can yield the highest mean AUC score of 83.0% on the Chest X-ray14 dataset and 94.6% on the Catheter dataset. According to the ablation study, we can obtain the following results: (1) The proposed MLOP embedding has a better performance than overlap patch (OP) embedding layer and non-overlap patch (N-OP) embedding layer that the mean AUC score is improved 0.6% and 0.4%, respectively. (2) Our demonstrate dynamic position feed forward can replace the traditional position mask which can learn the position information, and the mean AUC increased by 0.6%. (3) The mean AUC score by the designed MLA is more 0.2% and 0.6% than using the class token and calculating the mean scores of all tokens. The comprehensive experiments on two datasets demonstrate the effectiveness of the proposed method for multi-label chest X-ray image classification. Hence, our MXT can assist radiologists in diagnoses of lung diseases and check the placement of catheters, which can reduce the work pressure of medical staff.

Keywords Chest X-ray image · Multi-label classification · Transformer · Self-attention

Introduction

Globally, as of October 2021, there have been more than 236.60 million confirmed cases of COVID-19, including 4.83 million deaths, reported to WHO [1]. In October 2020, a new virus variant, Delta, was discovered in India for the first time which is more infectious than ordinary strains, and the infected are more likely to develop severe lung illness [2]. Under this unprecedented COVID-19 pandemic, many excellent works [3–9] based on deep learning were proposed for COVID-19 detection that can provide a fast and accurate diagnosis and severity assessment of viral pneumonia. Meanwhile, catheterization is a common treatment for patients, even more so now that millions of COVID-19

✉ Yu Zhu
zhuyu@ecust.edu.cn

✉ Dawei Yang
yang_dw@hotmail.com

[1] School of Information Science and Technology, East China University of Science and Technology, Shanghai 200237, People's Republic of China

[2] Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China

[3] Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai 200032, People's Republic of China

patients are in need of these tubes and lines to save their lives during the period of continued spread of the epidemic. The Chest X-ray (CXR) plays an important role in assisting clinical diagnosis which is one of the most common radiological examinations for screening catheters and diagnosis of many lung diseases.

However, improper placement of catheters may lead to serious complications. Nasogastric tube malpositioning into the airways has been reported in up to 3% of cases, and over 40% of these cases demonstrate complications [10]. In addition, airway tube malposition in adult patients intubated outside the operating room is seen in up to 25% of cases [11]. Doctors and nurses need frequently to check CXR images to diagnose the type of lung disease and determine if the catheters are placed correctly. Not only does this leave room for human error, but delays are also common as radiologists can be busy reporting other scans. Especially, CXR image classification is usually a multi-label problem in that a CXR image is usually labeled with one or more lung illnesses or types of catheter placement. Hence, automated multi-label CXR image classification has made significant progress to both diagnoses of lung diseases and check the placement of catheters.

In recent years, many studies in field of medical image analysis, such as classification [12, 13], image super-resolution [14, 15], and lesion segmentation or detection [16, 17], have advanced rapidly with the development of deep learning techniques. This study aims to study the problem of multi-label CXR image classification. Currently, many approaches [12, 18–22] for automated multi-label CXR image classification have been proposed to improve the classification performance and are mainly divided into three categories: (1) transfer learning approaches [12, 18]; (2) network innovation approaches [19, 20]; (3) attention guided approaches [21, 22]. However, although these works based on convolution neural network (CNN) have achieved excellent performance, they cannot fully meet the strict requirements of medical image analysis. One of the main reason is that the CNN-based method cannot learn global and long-range semantic information interaction, yielding to the locality of convolution operation [23].

Driven by great success of Transformer [24] in Natural Language Processing (NLP) domain, the researchers [25–31] have tried to introduce the Transformer into Computer vision (CV) domain. Furthermore, the self-attention of Transformer which can capture the short and long-range visual dependencies is the key to achieving outstanding results. Recently, pyramid vision Transformer (PVT) [31] which inherits the advantages from both CNN and Transformer is an alternative and useful backbone in CV domain. Motivated by the success of PVT, we proposed the MXT which is a new variant of PVT for multi-label chest X-ray image classification. The global computation of Transformer-based method leads to quadratic complexity concerning the number of tokens [28]. Therefore, we present downsampling spatial reduction attention to reducing the resource consumption of using Transformer. Meanwhile, we propose multi-layer overlap patch embedding layer to improve the performance of extracting features of each patch. The resolution of input performing through the original Transformer-based method usually remains the same due to the fixed size of position mask. To address this problem, we utilize dynamic position feed forward with zero paddings that can encode position to replace original positional mask. Based on the PVT backbone, we further add the class token Transformer block and multi-label attention to offer more effective processing of multi-label classification. Comparing with both CNN-based method (i.e., VGGNet [32], ResNet [33], DenseNet [34]) and Transformer-based method (i.e., ViT [26], PVT [31]), the comprehensive experiments on two datasets demonstrate that our MXT for multi-label chest X-ray image classification has a better performance.

The main contributions of this study are summarized as follows:

1. We propose a new variant Transformer-based model inspired by PVT for multi-label chest X-ray image classification, which can capture the short and long-range visual dependencies in CXR images.
2. Downsampling spatial reduction attention (DSRA) is presented to reduce the resource consumption of using Transformer. Meanwhile, multi-layer overlap (MLOL) patch embedding layer is to improve the performance of extracting features of each patch. Dynamic position feed forward with zero paddings is utilized to change the resolution of input flexibly.
3. We further add the class token Transformer block and multi-label attention to offer more effective processing of multi-label classification.

The overview of the study is organized as follows. "Related Work" introduces the related work in multi-label classification in deep neural networks and attention mechanisms. Then, the proposed MXT is described in "Proposed Method." Next, the comprehensive experiments and analysis are conducted in "Experimental Results and Analysis." At last, "Conclusion" concludes the whole work.

Related Work

Multi-label Classification in Deep Neural Network

Benefiting from the development of deep neural networks (DNN), classification tasks [33, 35–37] have made great progress. Traditional classification task studies the problem

where each example is represented by a single feature vector. Formally, the network is to learn a function $f : X \rightarrow Y$ from the training set. Here, X denotes the instance space and Y is the label space. However, real-world objects might be complicated and a single instance while associated with a set of labels simultaneously [38]. Hence, during the past decade, significant amount of researchers [36, 39] have paid attention to multi-label classification.

Multi-label CXR image classification as one of the most practical techniques in medical image processing plays an important role in assisting diagnosis [12] and [18] are both transfer learning approaches. Wang et al. [12] predict the presence of multiple disease on the Chest X-ray14 dataset via several classic CNN architecture (i.e., AlexNet [40], VGGNet [32], GoogLeNet [35], ResNet [33]). Meanwhile, Rajpurkar et al. [18] finetune a modified DenseNet [34] which replaces the last classification layer. Typically, network innovation approaches have expended considerable efforts in designing architectures. [19] presents a novel method termed a segmentation-based deep fusion network for thoracic disease classification in CXR images. [20] proposes a joint classification and regression method to jointly predict the disease progression and the conversion time. Inspired by the attention mechanism which has been successfully explored in the field of computer vision (CV), [21] proposes an attention branch, which learns discriminative attention maps and [22] designs category-wise residual attention learning.

Attention Mechanism

The attention mechanism can emphasize the regions that we focus on and suppress the irrelevant background regions through a set of weight coefficients learned by the network autonomously. At present, the attention mechanisms have many types of variants, such as channel attention, spatial attention, and self-attention that corresponding to different feature dimensions [41].

Channel Attention

The purpose of channel attention is to show the correlation between different channels, and automatically obtain the weights ($W \in R^{1\times1\times C}$) of feature maps ($M_C \in R^{H\times W\times C}$) in training process. Hence, the output of each feature channel is obtained by multiplying W and M_C, to strengthen the important features and suppress the unimportant features. SE-Net [42] constructs the channel importance weight function by fully connected layers. Inspired by the Inception-block [35] and SE block, SK-Net [43] introduces multiple convolution kernel branches to learn the attention of feature graphs at different scales so that the network can focus more on important scale features. In addition, ECA-Net [44] utilizes one-dimensional sparse convolution operation to optimize the full connection layer operation involved in SE block to greatly reduce the amount of parameters and maintain considerable performance. Above all, they are all representative channel attention mechanisms and achieve good results.

Spatial Attention

Spatial attention aims to improve the feature expression of important regions. In essence, it transforms the spatial information in the original feature map into another space and retains the important information by generating a learnable mask for each position. CBAM [45] proposed a SAM block that generates two feature maps $F_S \in R^{H\times W\times 2}$ with different information by global average pooling (GAP) and global max pooling (GMP). Then, F_S passes through a convolution layer and Sigmoid activation to generate the weight map to enhance the target area. In addition, Chen et al. proposed a novel spatial attention A^2-Net [46] that aggregates and propagates informative global features from the entire spatio-temporal space. Inspired by SE-Net, Guha Roy et al. [47] put forward three variants of scSE modules for image segmentation. These typical spatial attention mechanisms can enhance the specific target regions of interest in the feature maps and weaken the irrelevant background regions.

Self-attention

Self-attention was first proposed in Transformer [24] that can compute parallel for the machine translation task that has achieved the state-of-the-art performance. Driven by the great success of Transformer in NLP domain, the researchers [25–30] have tried to introduce the self-attention into Computer vision (CV) domain. Comparing the CNN-based methods [33, 35, 40, 48] that only focus on the information in the local receptive field, Transformer-based methods can pay attention to the global information in the whole images. ViT was proposed in [26] that directly introduced the Transformer encoder block into image recognition task using non-overlapping static image patches as tokens in NLP. However, the drawback of ViT is that it requires large-scale training datasets (i.e., JFT-300 M which contains 300 M images) although it achieved excellent speed accuracy. Hence, DeiT [30] utilized several training strategies and token-based distillation to reduce the size of training dataset (i.e., ImageNet-1 K dataset). Furthermore, Swin Transformer [28] propose W-MSA and SW-MSA block to calculate local self-attention and reduce resource consumption. However, it is complex to design the local windows and the information in each local window is not easy to interact. Pyramid vision Transformer (PVT) [31] which designs as a shrinking pyramid structure, inherits the advantages from both CNN and Transformer. Meanwhile, PVT utilizes simple

spatial-reduction attention (SRA) to reduce resource consumption and can achieve outstanding results. Inspired by PVT, we propose a new variant Transformer-based model to offer more effective processing of multi-label classification.

Proposed Method

Architecture Overview

The overall architecture of proposed MXT for multi-label CXR image classification is depicted in Fig. 1. Although ViT has achieved an outstanding result in classification, its feature map has only a single scale with low resolution that is not suitable for detecting targets with different scales. To address this problem, MXT introduces the shrinking pyramid structure into Transformer, which can reduce the resolution of the feature maps with the deepening of the network.

There are five stages in our MXT, and the first four stages share a similar architecture which consists of a multi-layer overlap (MLOL) patch embedding layer and a downsampling spatial reduction Transformer block. We further add the class token attention and multi-label attention in stage 5 to offer more effective processing of multi-label classification.

First, the MLOL patch embedding layer is applied to split the CXR image with the size of $H \times W \times 3$ into tokens and yields a feature map (FM) with size of $\frac{H}{4} \times \frac{W}{4} \times C_1$. Then, we flatten the feature map to a token map (TM) with size of $(\frac{H}{4} \times \frac{W}{4}) \times C_1$. The number of tokens is $(\frac{H}{4} \times \frac{W}{4})$, and the dimension of each token is C_1. After that, the token map passes through the DSR Transformer block, and the output feature is reshaped $\frac{H}{4} \times \frac{W}{4} \times C_1$. According to the same way, we use the feature maps from prior stages as input and obtain the following feature maps $\{MF_1, MF_2, MF_3\}$. Especially, we add a class token that is a learnable embedding vector to token map to predict the classification in stage 4. Significantly, we directly use the token map with the size of $(\frac{H}{32} \times \frac{W}{32}+1) \times C_4$ as the input of stage 5. At last, a class token Transformer and multi-label attention are applied to classify the CXR images.

Multi-layer Overlap Patch Embedding Layer

Different from original patch embedding layer of ViT [26] that directly crops the image into patches, we utilize multi-layer overlap (MLOL) patch embedding layer to tokenize

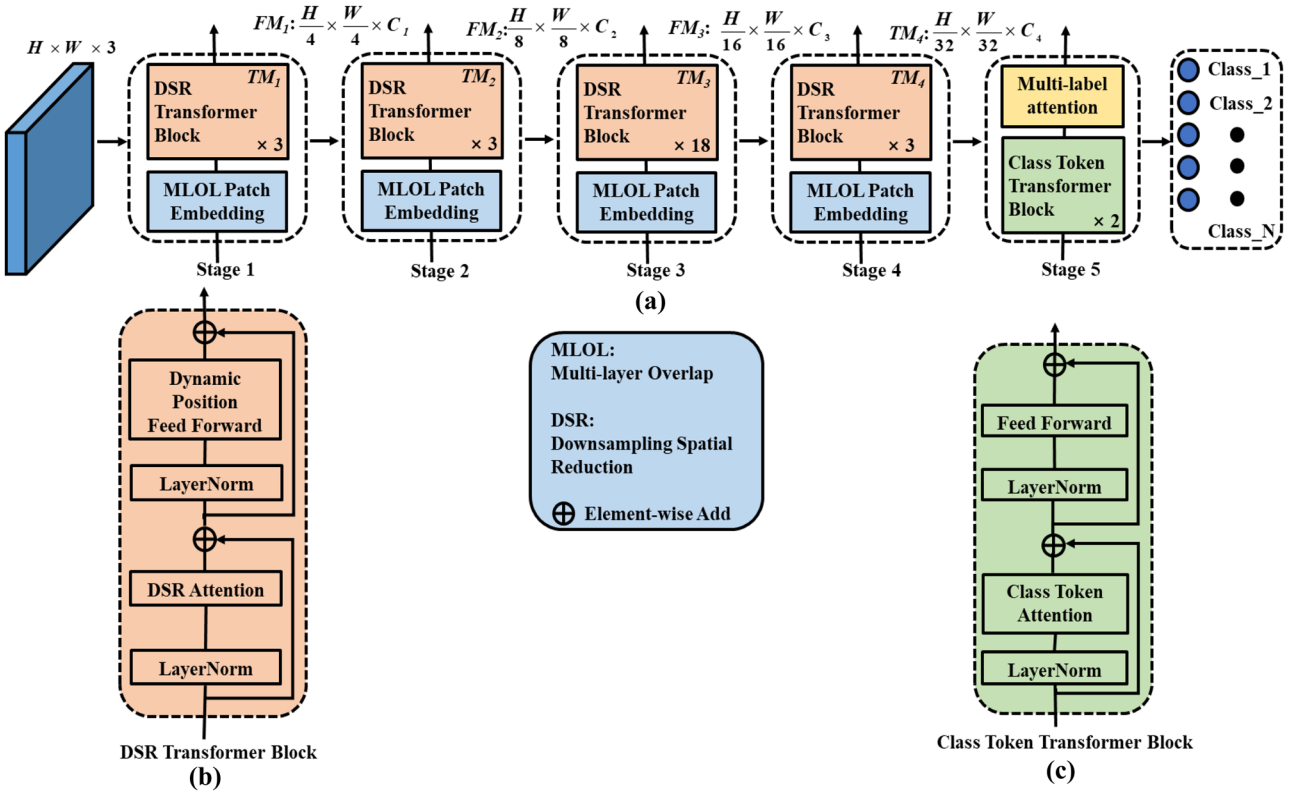


Fig. 1 The pipeline of the proposed MXT architecture. **a** Overall architecture, consisting of five stages. **b** Details of the downsampling spatial reduction Transformer block. **c** Construction of class token Transformer block

images by multi-layer convolutions. At first, we feed the given image of size $H \times W \times 3$ to a convolution with the stride of S, the kernel size of P as same as patch size, the padding size of S-1, and the kernel number of C. After that, the output feature map passes through another two convolution layers with the stride of 1, and the other settings are the same as the first layer. According to the hierarchical structure, the feature information between patches can be better fused. In addition, batch normalization and SiLU [49] activation function that looks like a continuous and "undershooting" version of the linear rectifier unit (ReLU [50]) which is shown in Eq. (1) are applied between two convolution layers. Finally, we flatten the feature map to a token map with size of $(\frac{H}{S} \times \frac{W}{S}) \times C$.

$$SiLU = x\frac{1}{1+e^{-x}} \tag{1}$$

Downsampling Spatial Reduction Transformer Block

As shown in Fig. 1b, downsampling spatial reduction (DSR) Transformer block in stage i has L_i encoder layers, and each layer is composed of DSR attention and dynamic position feed forward. For traditional multi-head attention (MHA) [24], the given token map passes through a linear projection and yields a query Q, a key K, and a value V as the input of MHA. Especially, to reduce the resource consumption of using Transformer, we reduce the spatial scale of K and V by a downsampling layer before MHA as illustrated in Fig. 2. Here, we reshape the token map $X \in \mathbb{R}^{(h\times w)\times C}$ to size of $h \times w \times C$ at first. Then, an average pooling with a pooling size of R is applied to generate a fixed feature map. To update the weight of the feature, the feature map passes through a convolution with kernel size of 1 and stride size of 1. Finally, the feature maps are reshaped to size of $R^2 \times C$ after performing layer normalization and SiLU activation function, as K and V. The attention operation is calculated as follows:

$$Attention(Q,K,V) = \text{Soft}\max(\frac{QK^T}{\sqrt{d_{head}}})V \tag{2}$$

Through the Eq. (2), we can clearly figure out that the computation costs of traditional MHA and downsampling spatial reduction attention (DSRA), which are as follows:

$$\Omega(MHA) = hwC^2 + (hw)^2C \tag{3}$$

$$\Omega(DSRA) = 2hwR^2C \tag{4}$$

The R^2 is far less than hw; therefore, DSRA can greatly reduce the resource consumption of using Transformer.

According to the research of [51], we can find that the zero-padding can imply the position information. Therefore, to adjust the resolution of input flexibly, we introduced dynamic position feed forward which is shown in Fig. 3, instead of adding the fixed position mask to model local spatial relationships. We first utilize a linear projection to add the dimensions of input. Then, the 2D token map $\mathbb{R}^{(h\times w)\times C}$ is reshaped to 3D $\mathbb{R}^{h\times w\times C}$ feature map before performing a convolution with kernel size of 3, and padding size of 1. Then, the 3D feature map is reshaped to 2D token map again. Finally, the second linear projection is applied to reduce the dimensions. According to dynamic position feed forward with zero padding, our MXT can adjust the resolution of input flexibly without the limit of fixed position mask.

Method of Calculating the Classification Scores

Class Token Transformer Block

As shown in Fig. 1c, class token Transformer block is introduced into stage 5 that has two encoder layers, and each layer is composed of class token attention and feed forward. Simulate to the typical Transformer, we concatenate a class token that is a learnable embedding vector to token map to predict the classification in stage 4. However, it is difficult for the class token to focus on attention map and perform

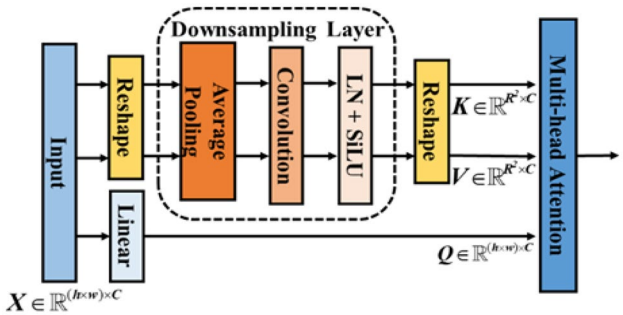


Fig. 2 The pipeline of downsampling spatial reduction attention

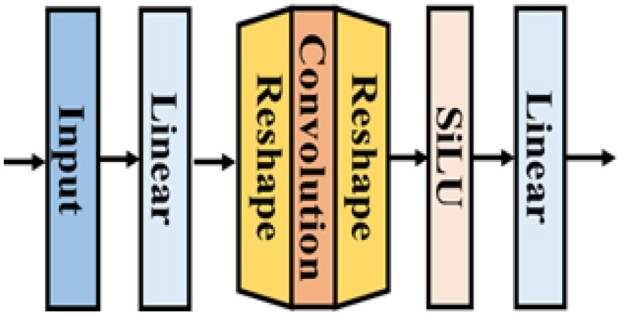


Fig. 3 Dynamic position feed forward

```
#Multi-label attention pseudo code

class MLA(nn.Module):
#x: Token map, shape: B, N, C

    def __init__(self, channel=512 , num_class=14, la=0.2):
        super().__init__()
        self.la=la
        self.fc=nn.Linear(channel, num_class)

    def  forward(self, x):
        x=self.fc(x)
        class_token=x[:, 0]
        score_max=torch.max(x,dim=1)[0]
        MLA_score=class_token+self.la*score_max
        return MLA_score
```

Fig. 4 Multi-label attention pseudo code

the classification at the same time. Therefore, the class token is separated from the token map in stage 5 and then passes through a linear projection to yield a query $Q \in \mathbb{R}^{1\times C}$, a key $K \in \mathbb{R}^{1\times C}$, and a value $V \in \mathbb{R}^{1\times C}$. Then, MHA is applied to Q, K, and V. After that, a feed forward layer is utilized to choose the character. In this stage, only the class token is updated.

Multi-label Attention

Multi-label attention (MLA) pseudo code is shown in Fig. 4. We considered combining the class token scores and max pooling scores as the final predicted score, and *la* is a hyper-parameter to balance them. On the one hand, max pooling can find the maximum value among all spatial locations for each category. Hence, it can be considered as a class-specific attention mechanism. On the other hand, MLA enables our model to focus on the classification scores of different object categories at different locations. Compared with the traditional classification network, MLA is more suitable for the multi-label classification task and detailed ablation experiments will be carried out in "Ablation Study."

Detailed Setting of MXT

According to the typically shrinking pyramid structure, MXT can extract the target information of different scales in CXR image more efficiently. The detailed setting of MXT, simulating the CNN-based ResNet [33], is shown in Table 1. Here, S_i and P_i are respectively the stride and patch size of the MLOL in Stage i. C_i stands the channel number of the output of Stage i. Meanwhile, R_i is the adaptive average pooling size of the DSR Transformer in Stage i. K_i and $padding_i$ represent the kernel size and padding size of dynamic position feed forward in Stage i. L_i is the number of DSR Transformer layers in Stage i.

Implements for Pre-processing CXR Images

The size of CXR images is 1024×1024. During training, we resized the input images to 384×384 to fairly compare with existing methods. Then, center cropping and random horizontal flip were employed to augment images. During testing, the test images were also resized to 384×384, and performed the same data augmentation.

Experimental Results and Analysis

Dataset

We evaluate our MXT on a large-scale CXR dataset, Chest X-ray14 [52], released by National Institutes of Health Clinical Center (NIH). The dataset consists of 30,805 patients and 112,120 frontal-view X-ray images with 14 disease pathologies, and each image is labeled one or more pathologies. We illustrate some example CXR images and corresponding disease pathologies in Fig. 5. Especially, we follow [12] to split the dataset, including three subgroups (70% training, 10% validation, and 20% testing). Each image is labeled with $L = \{I_1, I_2, ..., I_C\}$, and C is 14 in dataset. Each element of I is set as 0 for absence and 1 for presence.

Furthermore, the Royal Australian and New Zealand College of Radiologists (RANZCR) which is a not-for-profit

Table 1 The detailed setting of MXT. The design simulates the CNN-based ResNet [33] that reduces the resolution of the feature maps and increases the hidden dimension with the deepening of the network

	Output size	Layer name	Hyper-parameters
Stage 1	$\frac{H}{4}\times\frac{W}{4}\times C_1$	MLOL	$S_1=4, P_1=7, C_1=64$
		DSR Transformer	$R_1=7, K_1=3, Padding_1=1, L_1=3$
Stage 2	$\frac{H}{8}\times\frac{W}{8}\times C_2$	MLOL	$S_2=2, P_2=3, C_2=128$
		DSR Transformer	$R_2=7, K_2=3, Padding_2=1, L_2=3$
Stage 3	$\frac{H}{16}\times\frac{W}{16}\times C_3$	MLOL	$S_3=2, P_3=3, C_3=256$
		DSR Transformer	$R_3=7, K_3=3, Padding_3=1, L_3=18$
Stage 4	$(\frac{H}{32}\times\frac{W}{32}+1)\times C_4$	MLOL	$S_4=2, P_4=3, C_4=512$
		DSR Transformer	$R_4=7, K_4=3, Padding_4=1, L_4=3$

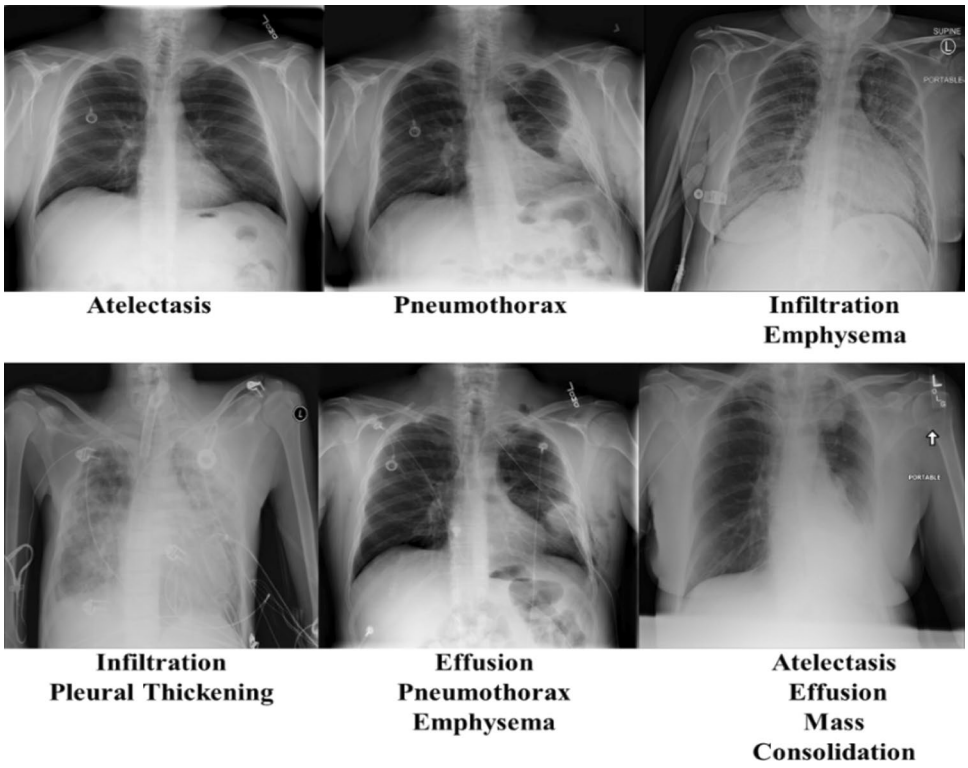


Fig. 5 Example CXR images and their corresponding disease pathologies

professional organization for clinical radiologists and radiation oncologists in Australia, New Zealand, and Singapore organized *Catheter and Line Position Challenge* [53] on Kaggle. There are 30,083 CXR images with 11 types of catheter placement, and each image may have one or more types. They are also selected from the publicly available ChestXRay14 dataset. Some types of catheter placement are shown in Fig. 6. We randomly split the dataset into training data (80%) and testing data (20%). Each CXR image is also labeled with one-hot vector. The quantities of each type of catheter placement is shown in Fig. 7, we can clearly find that the dataset is highly imbalanced and this is a great challenge for our MXT.

Implementation Details

Furthermore, the proposed MXT framework is implemented by using Python 3.6 and Pytorch 1.7.0 that runs on 4 Nvidia 1080Ti GPU with 12 GB memory. The weights pre-trained on ImageNet [54] are used to initialize the model parameters. During the training period, the mini-batch size is 36 and the initial learning rate is set to $1e-4$ and decreases

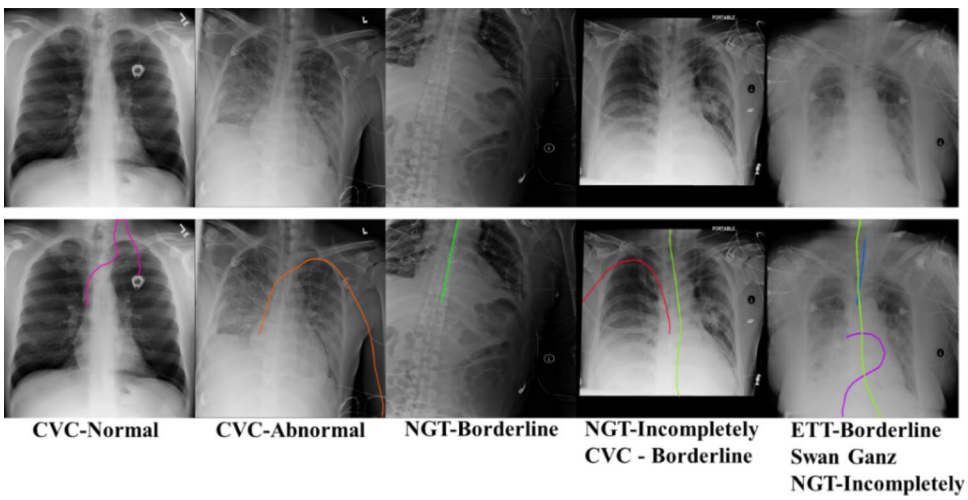
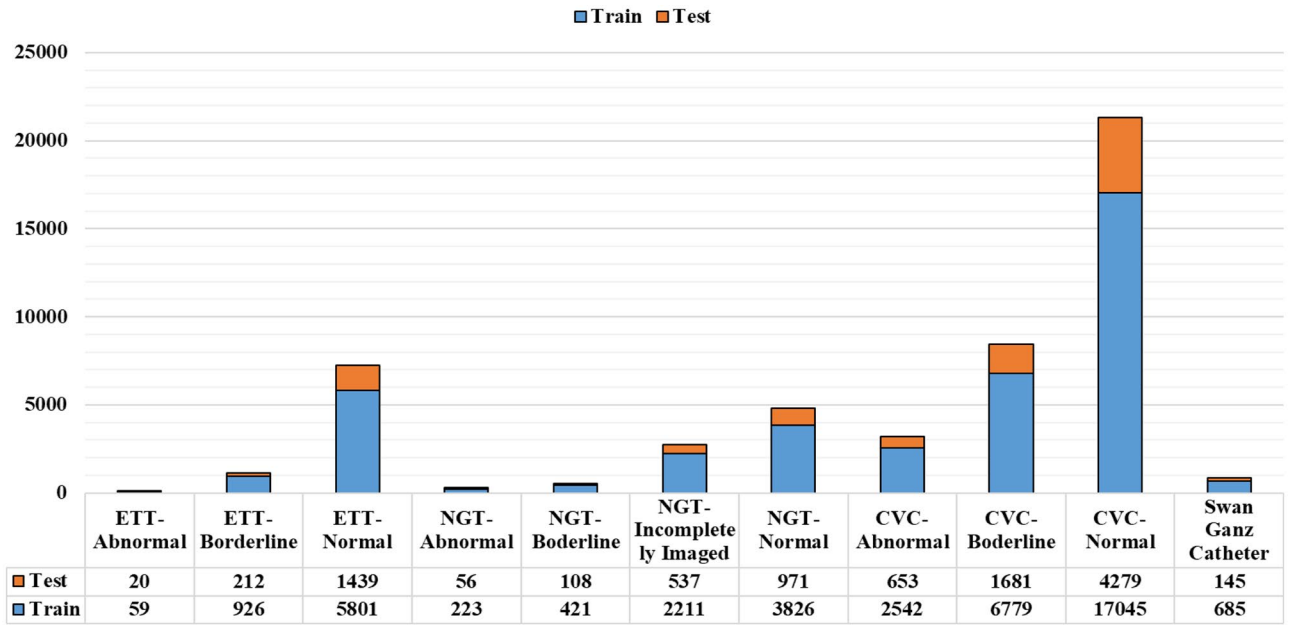


Fig. 6 Example CXR images and their corresponding types of catheter placement



	ETT-Abnormal	ETT-Borderline	ETT-Normal	NGT-Abnormal	NGT-Boderline	NGT-Incompletely Imaged	NGT-Normal	CVC-Abnormal	CVC-Boderline	CVC-Normal	Swan Ganz Catheter
Test	20	212	1439	56	108	537	971	653	1681	4279	145
Train	59	926	5801	223	421	2211	3826	2542	6779	17045	685

Fig. 7 The quantities of each type of catheter placement. The red bar is training data, and the blue one is test data. CVC means central venous catheter, and NGT is nasogastric tube, and ETT stands for endotracheal tube

following the cosine schedule [55] for 20 epochs. In addition, the popular AdamW [56] optimizer with momentum 0.9 and weight decay $1e-3$ is used to optimize our model for back propagation. Especially, to deal with the problem of class imbalance, we utilized the weighted cross-entropy loss (W-CELoss) as our loss function

$$W - CELoss = -w_P y_i \log(f(x_i)) - w_N(1 - y_i)\log(f(x_i)) \quad (5)$$

where $w_p = \frac{P+N+1}{P+1}$, and $w_N = \frac{P+N+1}{N+1}$, that P and N are the total numbers of positive and negative labels in a mini-batch of image input, respectively.

Evaluation Metrics of Multi-label Classification

Multi-label evaluation metrics can be generally categorized into two groups, i.e., label-based metrics and example-based metrics [38]. Label-based metrics evaluate the learning system's performance on each class label separately at first and then returning the macro- or micro-averaged value across all class labels. Unlike the above label-based metrics, example-based metrics evaluate the learning performance of system on each test example separately and then returning the mean value across the test set. Following the previous works [12,

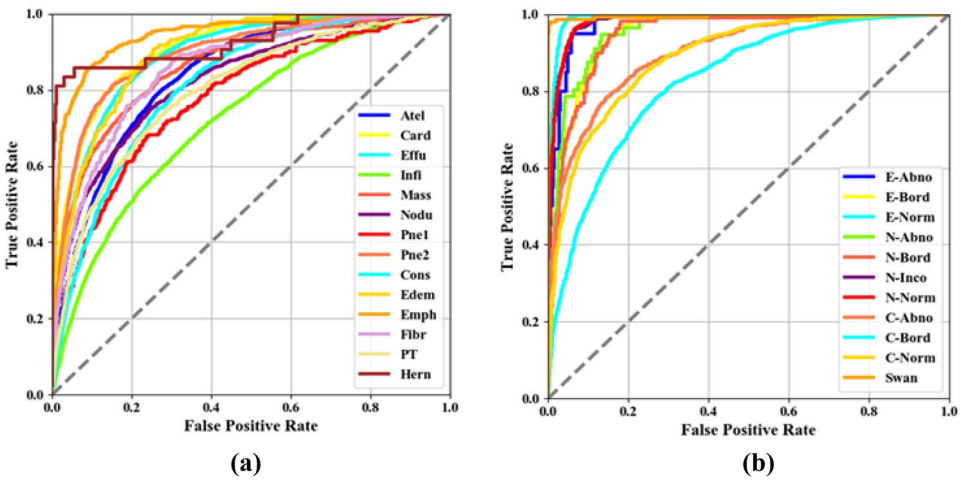


Fig. 8 ROC curves of the proposed MXT framework on the ChestX-ray14 (**a**) and Catheter (**b**) datasets. The corresponding AUC scores are given in Tables 2 and 3

Table 2 Comparison results of various methods on ChestX-ray14 dataset. We illustrate the AUC score (%) of each disease pathology and the average AUC scores (%) across the 14 classes. The highest scores are shown in bold

Method	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
Model A [12]	70.0	81.0	75.9	66.1	69.3	66.9	65.8	79.9	70.3	80.5	83.3	78.6	68.4	87.2	74.5
Model B [20]	73.3	85.6	80.6	67.3	71.8	77.7	68.4	80.5	71.1	80.6	84.2	74.3	72.4	77.5	76.1
Model C [21]	75.1	87.1	81.2	68.1	79.9	71.5	69.4	82.5	74.2	83.5	84.3	80.4	74.6	90.2	78.8
Model D [18]	76.3	87.1	82.3	69.4	82.0	74.9	71.8	86.1	73.9	84.0	92.2	82.9	77.4	89.0	80.7
Model E [19]	78.1	88.5	83.2	70.0	81.5	76.5	71.9	86.6	74.3	84.2	92.1	83.5	79.1	91.1	81.5
Model F [22]	78.1	88.0	82.9	70.2	83.4	77.3	72.9	85.7	75.4	**85.0**	**90.8**	83.0	77.8	**91.7**	81.6
ViT [26]	77.3	88.8	82.0	70.8	81.8	75.6	73.2	86.2	75.3	85.3	87.9	82.4	76.1	88.2	80.8
PVT [31]	78.6	88.1	83.3	70.7	84.9	80.6	74.1	87.6	**75.9**	84.7	89.7	83.7	77.9	89.8	82.1
Our MXT	**79.8**	**89.6**	**84.2**	**71.9**	**85.6**	**80.9**	**75.8**	**87.9**	**75.9**	84.9	90.6	**84.7**	**80.0**	91.3	**83.0**

13, 22], we used the area under curve (AUC) that is the receiver operating characteristic curve (ROC) [57] which is label-based metric as our main evaluation metric to compare with some state-of-the-art methods. The abscissa of ROC curve is the false positive rate, and the ordinate is the true positive rate. Meanwhile, the hamming loss (HL),the one-error (OE), and the label ranking average precision (LRAP) [38] are considered as additional evaluation metrics in this study, which are typical example-based metrics for multi-label classification. The hamming loss evaluates the fraction of misclassified instance-label pairs that a relevant label is missed or an irrelevant is predicted

$$HL = \frac{1}{N}\sum_{i=1}^{N}\frac{XOR(L_i, Y_i)}{C} \tag{6}$$

Here, N is the number of samples, and C represents the number of categories, while L_i is a one-hot vector label and Y_i stands for the one-hot vector prediction that we set 0.5 as the threshold for the MXT output. As shown in Eq. (7), the OE evaluates the fraction of examples whose top-ranked label is not in the relevant label set.

$$OE = \frac{1}{N}\sum_{i=1}^{N}\arg\max Y_i \notin L_i \tag{7}$$

LRAP is the average over each ground truth label assigned to each sample, of the ratio of true vs. total labels with a lower score.

$$LRAP = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|L_i|}\sum_{l\in L_i}\frac{\left|\left\{l' | rank_f(y, l') \le rank_f(y_i, l), l' \in L_i\right\}\right|}{rank_f(y_i, L)} \tag{8}$$

Comparison with State-of-the-Art Methods

On the Chest X-ray14 dataset, we evaluated the proposed MXT model against six recent CNN-based method and two Transformer-based method (ViT [26] and PVT [31]). The CNN-based methods have been introduced in detail in "Multi-label Classification in Deep Neural Network" and are abbreviated as Models A [12], B [20], C [21], D [18], E [19], and F [22], respectively. On the Catheter dataset, we utilize the CNN-based backbone (i.e., EffNet [37], ResNet50 [33], and Densenet121 [34]) and Transformer-based method (ViT [26] and PVT [31]) to evaluate our MXT.

The ROC Curves and AUC Score

The ROC curves and AUC score are first utilized to evaluate our MXT on the Chest X-ray14 dataset and

Table 3 Comparison results of various methods on Catheter dataset. We illustrate the AUC score (%) of each class and the average AUC scores (%) across the 11 types of catheter placement. The highest scores are shown in bold

Method	ETT			NGT				CVC			Swan	Mean
	Abno	Bord	Norm	Abno	Bord	Inco	Norm	Abno	Bord	Norm		
EffNet [37]	92.4	93.6	98.8	91.3	91.0	96.4	96.6	83.2	76.4	82.6	99.5	91.0
ResNet50 [33]	96.5	95.0	99.0	92.1	92.8	97.3	97.7	86.0	80.0	85.7	**99.7**	92.8
Densenet121 [34]	97.5	**95.1**	**99.1**	94.5	93.6	97.4	97.9	88.3	80.3	86.6	**99.7**	93.6
ViT [26]	95.6	93.8	98.7	91.6	92.3	97.0	97.5	85.6	79.8	85.0	99.6	92.3
PVT [31]	97.8	95.0	99.0	93.5	93.9	97.9	**98.6**	88.6	81.2	88.4	99.6	93.9
Our MXT	**97.9**	**95.1**	**99.1**	**95.8**	**94.6**	**98.3**	98.5	**90.1**	**82.9**	**89.0**	99.6	**94.6**

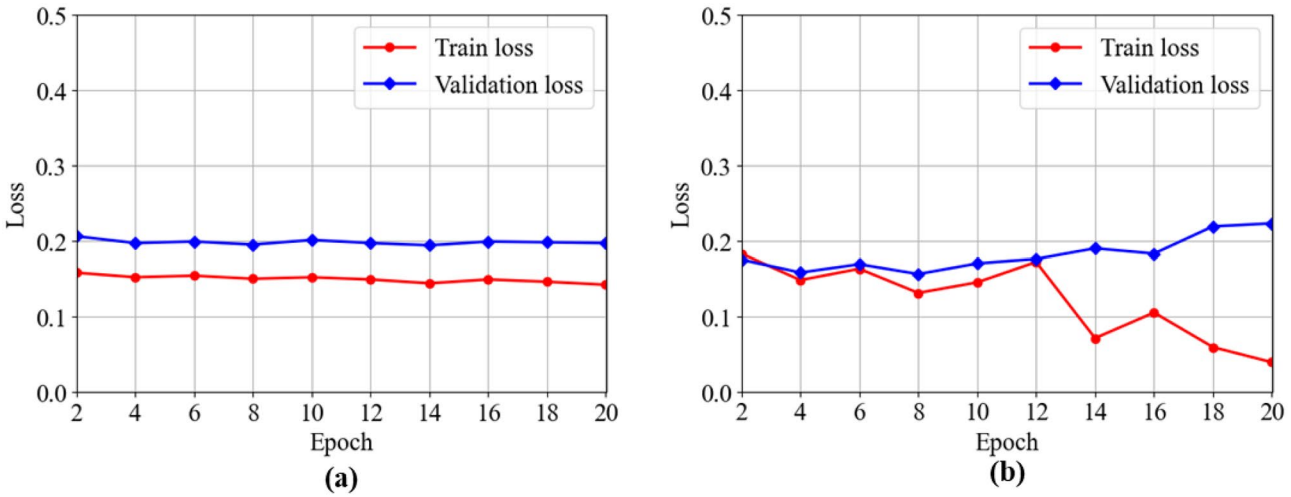


Fig. 9 Training and validation loss function with respect to epochs of the proposed MXT framework on the ChestX-ray14 (**a**) and Catheter (**b**) datasets, respectively

the Catheter dataset. The ROC curves of the proposed MXT framework over the 14 pathologies are illustrated in Fig. 8a, while the 11 types of catheter placement are shown in Fig. 8b. Moreover, the comparative results of our MXT method and the state-of-the-art baselines are presented in Tables 2 and 3. Based on the above results, we can gain the following observations: (1) Our MXT can yield the highest mean AUC score of 83.0% on the Chest X-ray14 dataset and 94.6% on the Catheter dataset, which demonstrates the efficacy of our method. (2) Subject to the experimental conditions, we set the batch size as 32 when finetuning ViT model. On the contrary, [26] set the batch size like 512 to transfer learning. Hence, the mean AUC score is only 80.8% on the Chest X-ray14 dataset that is 0.7% lower than Model D and 0.8% lower than Model F. Meanwhile, the mean AUC score of 92.3% on the Catheter dataset that is 1.3% lower than Densenet121. (3) According to shrink pyramid structure, our MXT can use a small batch size (i.e., 32), to achieve good experimental results. The mean AUC score is 8.5%, 6.9%, 4.2%, 2.3%, 1.5%, and 1.4% more than Models A, B, C, D, E, and F on the Chest X-ray14 dataset, respectively. Meanwhile, the mean AUC score on the Catheter dataset is 3.6%, 1.8%, and 1.0% more than EffNet, ResNet50, and Densenet121, respectively. (4) With architectural components and choices, our MXT can improve the performance that the mean AUC score is 0.9% and 0.7% more than PVT on Chest X-ray14 dataset and Catheter dataset, respectively.

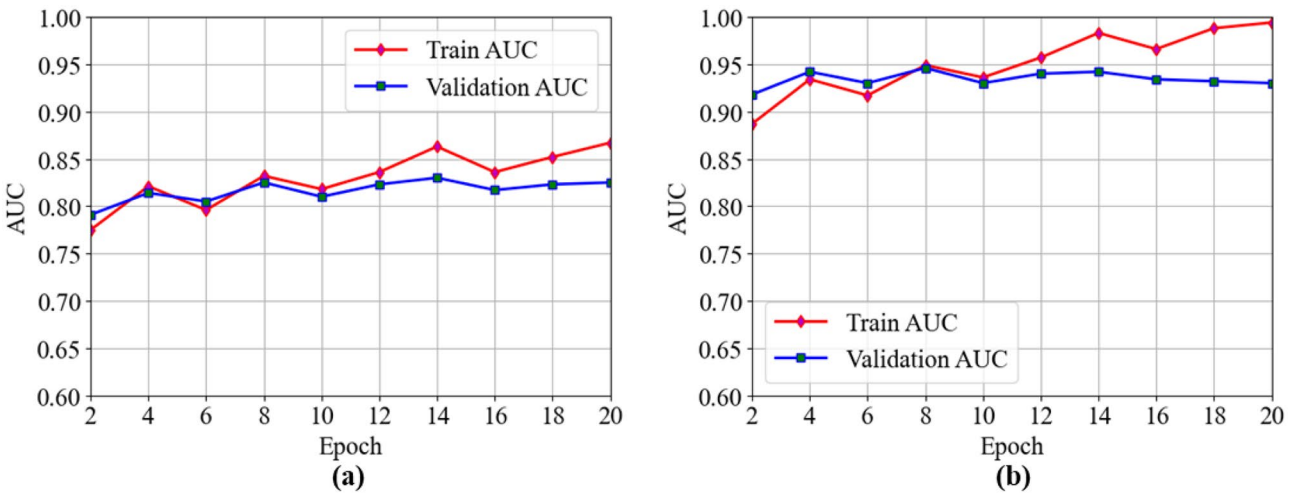


Fig. 10 Training and validation AUC scores with respect to epochs of the proposed MXT framework on the ChestX-ray14 (**a**) and Catheter (**b**) datasets, respectively

Table 4 Comparison of other evaluation metrics on the ChestX-ray14 dataset. The best performances are shown in bold

Method	HL	LRAP(%)
Model A [12]	0.091	72.4
Model D [18]	0.083	78.5
ViT [26]	0.077	78.6
PVT [31]	0.074	80.5
Our MXT	**0.069**	**81.0**

Table 5 Comparison of other evaluation metrics on the Catheter dataset. The best performances are shown in bold

Method	OE	HL	LRAP(%)
EffNet [37]	0.198	0.077	86.3
ResNet50 [33]	0.174	0.068	88.2
Densenet121 [34]	0.170	0.066	88.3
ViT [26]	0.172	0.068	88.2
PVT [31]	0.156	0.062	89.3
Our MXT	**0.145**	**0.059**	**90.1**

The 14 pathologies are atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia, respectively.

The 11 types of catheter placement are ETT-abnormal, ETT-borderline, ETT-normal, NGT-abnormal, NGT-borderline, NGT-incompletely Imaged, NGT-normal, CVC-abnormal, CVC-borderline, CVC-normal, and Swan Ganz Catheter Present, respectively. Here, ETT means endotracheal tube, NGT is nasogastric tube, and CVC stands central venous catheter.

Loss Function and AUC Scores with Respect to Epochs

Figure 9a and b illustrate the training and validation loss function with respect to epochs of our MXT framework on the two datasets, while the AUC scores with respect to

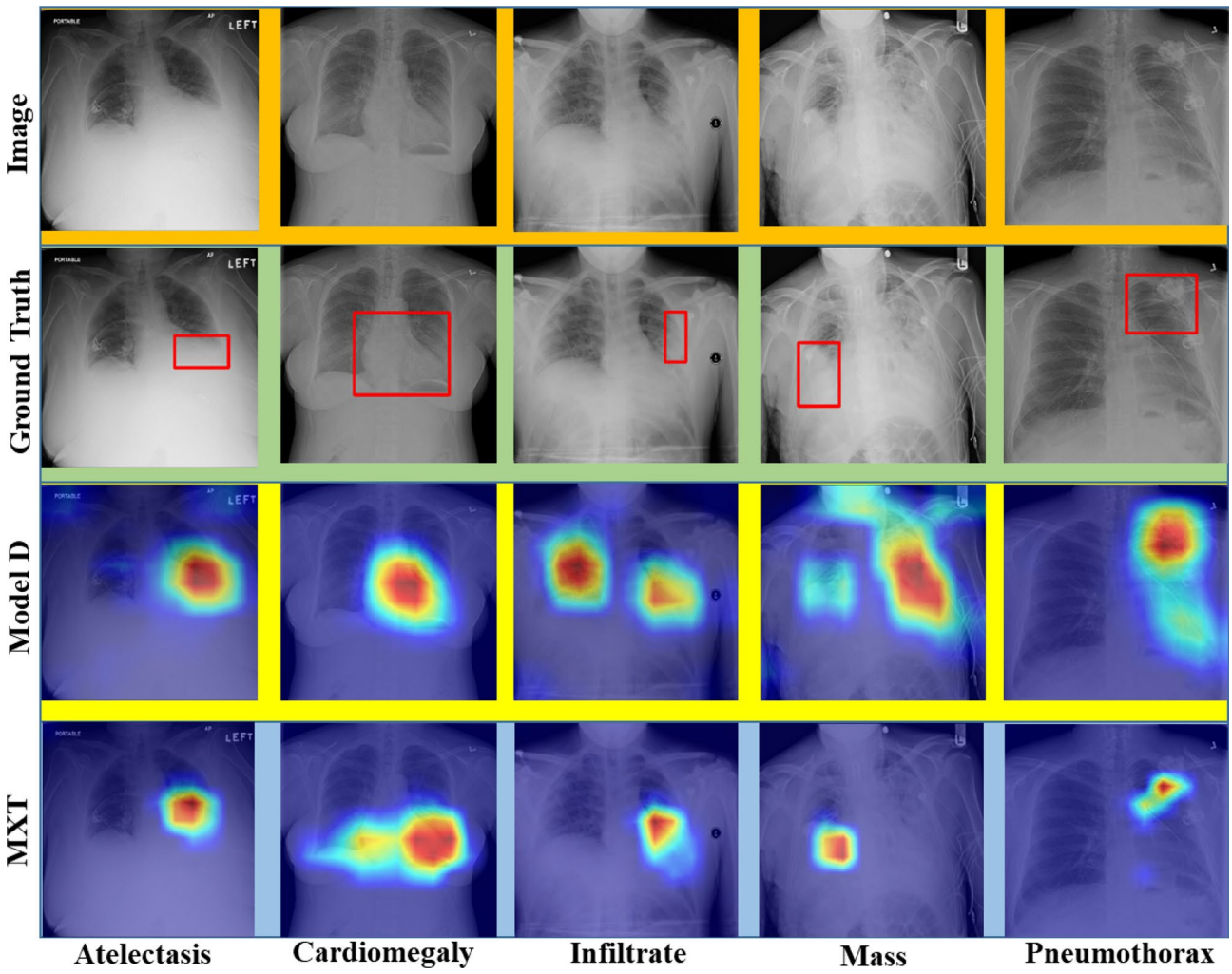


Fig. 11 Comparison of the localization of lesion regions. The first line is the original image, and the second line is added to the ground truth bounding boxes. The heatmaps of Model D and our MXT model are respectively illustrated in the third and the fourth line

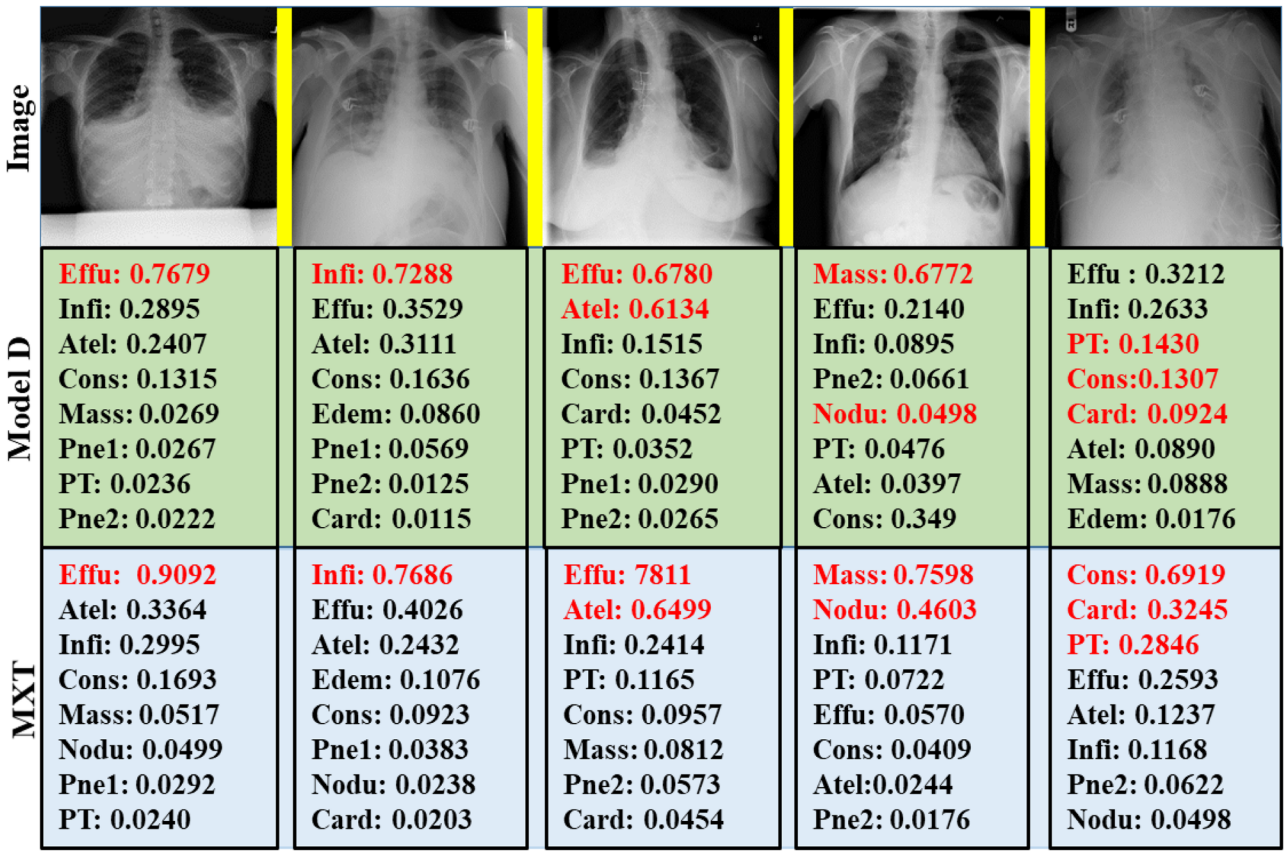


Fig. 12 Qualitative results on Chest X-ray14 dataset. The first line is the original image. The second line is the prediction results of Model D, and the third line is MXT. The ground truth categories are highlighted in red

epochs are shown in Fig. 10. On the basis of the curves, we can find that the highest validation AUC scores on the ChestX-ray14 occurred at 14th epoch, while occurring at 8th epoch on the Catheter.

Comparison of Other Evaluation Metrics

Moreover, we utilize the typical multi-label classification metrics, such as HL, the OE, and the LRAP to further evaluate our MXT method. From Tables 4 and 5, we can find the following observations: (1) Our MXT achieves the lowest HL (0.069) and the highest LRAP (81.0%) on the Chest X-ray14 dataset. The LRAP is 8.6%, 2.5%, and 2.4% higher than Model A, Model D, and ViT. The result means that our MXT can give a better rank for the label related to each sample. (2) Compared with the CNN-based method, such as EffNet, ResNet50, and Densenet121 and Transformer-based method (i.e., ViT and PVT), our MXT gain the most outstanding performance on the Catheter dataset, evaluated by OE, HL, and LRAP. Here, the OE is 0.053, 0.029, and 0.025 lower than EffNet, ResNet50, and Densenet121. It means that the top-ranked label generated by our MXT is more likely to be a truth label. (3) Compared with PVT, our MXT can also perform better that the HL decreases by 0.005 and 0.003, and the LRAP increases by 0.5% and 0.8% on the two datasets, respectively.

Localization of Lesion Regions

We further utilize the gradient-weighted class activation map (Grad-CAM) to generate the heatmaps of the CXR images that can approximately visualize the indicative attention areas. Examples of heatmaps generated from the Model D [18] which is a representative CNN-based method for multi-label CXR image classification and our MXT are shown in Fig. 11, and we can find that the highlighted regions on the heatmaps generated from our MXT are pretty closer to ground-truth, comparing to the Model D. The heatmaps generated from the Model D are discrete and blurry (i.e., CXR image with Atelectasis, Pneumothorax), and location of the lesion regions is in error (i.e., CXR image with Infiltrate, Mass). On the contrary, our MXT can capture the lesion regions more accurately and has a larger feeling field (i.e., CXR image with Cardiomegaly), benefiting from the self-attention which can pay attention to the global information in the whole CXR image.

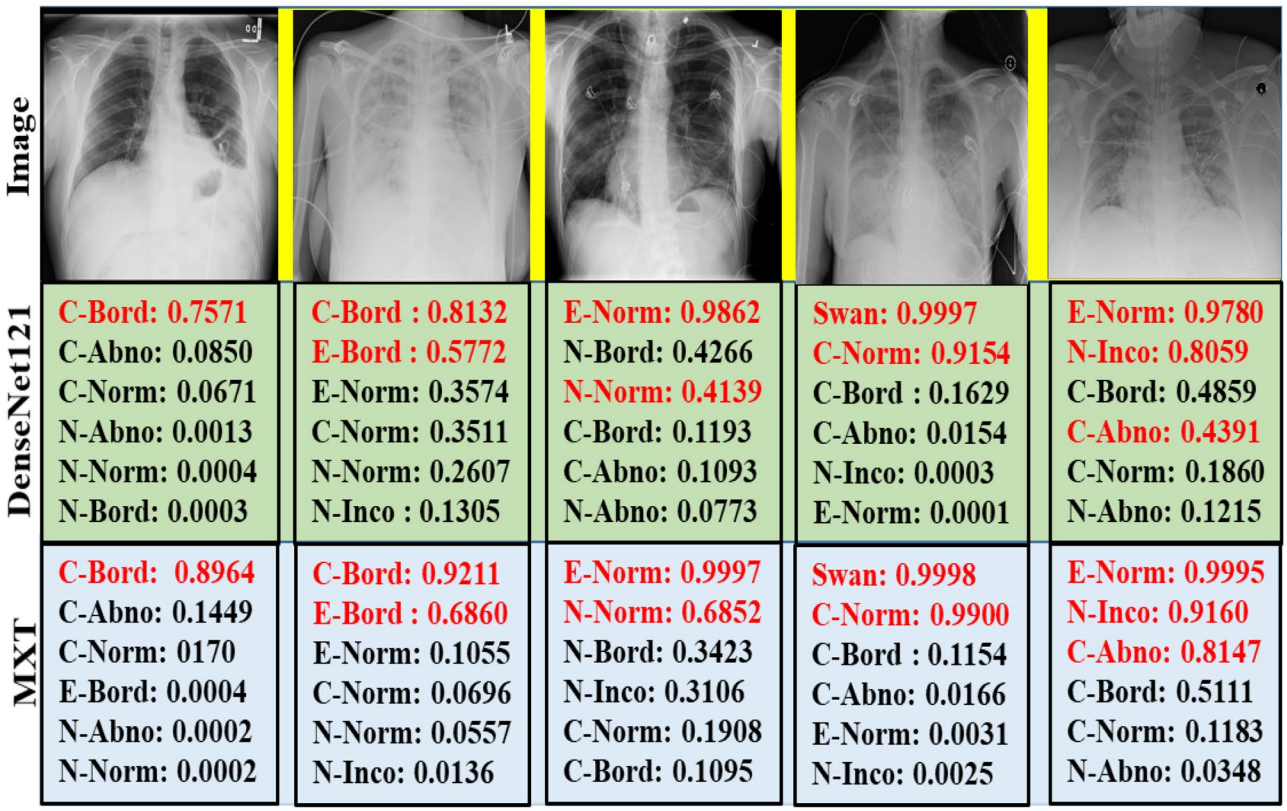


Fig. 13 Qualitative results on Catheter dataset. The first line is the original image. The second line is the prediction results of DenseNet121, and the third line is MXT. The ground truth categories are highlighted in red

Qualitative Results of Multi-label Classification

Figures 12 and 13 illustrate the intuitive presentations of our classification results on the Chest X-ray14 dataset and the Catheter dataset, respectively. Especially, the top-8 predicted categories and their corresponding prediction scores are presented in detail. On the basis of Figs. 12 and 13, we can clearly find that our MXT can obtain more accurate and reliable classification results. For example, in columns 1, 2, and 3 of Fig. 12, the scores of ground truth pathologies are higher than Model D, and columns 4 and 5 have the wrong prediction results. Columns 3 and 5 of Fig. 13 generated by DenseNet121 get the wrong prediction ranking, while our MXT correctly orders and gets higher scores of ground truth catheter types.

Ablation Study

We design ablation studies to explore the impact of our architectural components and choices on the final performance. All the ablation experiments use the same test data from Catheter dataset.

Multi-layer Overlap Patch Embedding Layer

First, to evaluate the contribution of the proposed MLOP embedding layer to our model, we utilize the traditional overlap patch (OP) embedding layer and non-overlap patch (N-OP) embedding layer to do comparative experiments, as shown in Table 6. From Table 6, we can find that MLOP embedding has a better performance than the OP embedding layer and N-OP embedding layer that the mean AUC score is improved 0.6% and 0.4%, respectively.

Dynamic Position Feeds Forward

Then, we explore the influence of dynamic position feed forward layer. The results are shown in Table 7, and demonstrate dynamic position feed forward can replace the traditional position mask which can learn the position information, and the mean AUC increased by 0.6%. On the contrary, the mean AUC decreases by 1.4%, and the HL increases by 0.006 while removing dynamic position feed forward and position mask. Further, when using both of them as Table 7B, it slightly drops 0.1%.

Table 6 Comparison of the influence of different patch embedding layers. The best performances are shown in bold

	OP embedding layer	N-OP embedding layer	MLOP embedding layer
AUC (%)	94.0	94.2	**94.6**
HL	0.062	0.061	**0.059**

Table 7 Ablations on dynamic position feed forward. The best performances are shown in bold

	Dynamic position feed forward	Position mask	AUC (%)	HL
A			93.2	0.065
B	✓	✓	94.5	**0.059**
C		✓	94.0	0.061
D	✓		**94.6**	**0.059**

Table 8 Comparison of the method of calculating the classification scores. The best performances are shown in bold

	Class token	All tokens	MLA (la = 0.1)	MLA (la = 0.2)	MLA (la = 0.3)
AUC (%)	94.4	94.0	94.3	**94.6**	94.5
OL	0.060	0.061	0.060	**0.059**	**0.059**

Multi-label Attention

Finally, we study the influence of multi-label attention (MLA) on the CXR image classification. In stage 5, we directly use the scores of the class token, or calculate the mean scores of all tokens as comparative experiments. According to Table 8, we can find that our MLA with la = 0.2 can gain the best performance. The mean AUC increases by 0.2% and 0.6% than using the class token and calculating the mean scores of all tokens.

Conclusion

In this research work, we propose a new variant of pyramid vision Transformer (MXT) for multi-label chest X-ray image classification. According to self-attention, our MXT can capture both short and long-range visual information in CXR images. We use downsampling spatial reduction attention to reduce the resource consumption of using Transformer. Meanwhile, multi-layer overlap patch embedding is utilized to tokenize images and dynamic position feed forward with zero paddings can encode position instead of adding a positional mask. Furthermore, class token Transformer block and multi-label attention are utilized to offer more effective processing of multi-label classification. Extensive experiments on two datasets illustrate that our MXT is efficient for multi-label chest X-ray image classification. In the future, we will focus on introducing the Transformer into medical image processing, such as image segmentation and disease detection. We hope that these study work can assist radiologists in diagnoses of lung diseases and check the placement of catheters, which can reduce the work pressure of medical staff.

Funding The authors received financial support from National Scientific Foundation of China (82170110), Shanghai Pujiang Program (20PJ1402400), Zhongshan Hospital Clinical Research Foundation (2019ZSGG15), and Science and Technology Commission of Shanghai Municipality (20DZ2254400, 21DZ2200600, 20DZ2261200).

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest The authors declare no competing interests.

References

1. WHO. WHO Coronavirus (COVID-19) Dashboard. 2021. https://covid19.who.int/.
2. Xia F, Yang X, Cheke RA, Xiao Y. Quantifying competitive advantages of mutant strains in a population involving importation and mass vaccination rollout. Infectious Disease Modelling. 2021;6:988–96.
3. Paul A, Basu A, Mahmud M, Kaiser MS, Sarkar R. Inverted bell-curve-based ensemble of deep learning models for detection of COVID-19 from chest X-rays. Neural Comput Applic. 2022:1–15.
4. Prakash N, Murugappan M, Hemalakshmi G, Jayalakshmi M, Mahmud M. Deep transfer learning for COVID-19 detection and infection localization with superpixel based segmentation. Sustainable Cities Society & Natural Resources. 2021;75: 103252.
5. Kumar S, Viral R, Deep V, Sharma P, Kumar, M, Mahmud M, et al. Forecasting major impacts of COVID-19 pandemic on country-driven sectors: challenges, lessons, and future roadmap. Personal Ubiquit Comput. 2021:1–24.
6. Gomes JC, Barbosa VAdF, Santana MA, Bandeira J, Valença MJS, de Souza RE, et al. IKONOS: an intelligent tool to support diagnosis of COVID-19 by texture analysis of X-ray images. Research on Biomedical Engineering. 2020:1–14.
7. Ismael AM, Şengür A. The investigation of multiresolution approaches for chest X-ray image based COVID-19 detection. Health Information Science Systems. 2020;8(1):1–11.
8. Gomes JC, Masood AI, Silva LHdS, da Cruz Ferreira JRB, Freire Junior AA, Rocha ALdS, et al. Covid-19 diagnosis by combining RT-PCR and pseudo-convolutional machines to characterize virus sequences. Sci Rep. 2021;11(1):1–28.
9. Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Syst Appl. 2021;164: 114054.

10. Sorokin R, Gottlieb JE. Enhancing patient safety during feeding-tube insertion: a review of more than 2000 insertions. J Parenter Enter Nutr. 2006;30(5):440–5.
11. Lotano R, Gerber D, Aseron C, Santarelli R, Pratter M. Utility of postintubation chest radiographs in the intensive care unit. Crit Care. 2000;4(1):1–4.
12. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Conference on Computer Vision and Pattern Recognition (CVPR). 2017; pp. 2097–2106.
13. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2019; pp. 590–597.
14. Mahapatra, D, Bozorgtabar, B, Garnavi, R, Graphics. Image super-resolution using progressive generative adversarial networks for medical image analysis. Comput Med Imaging Graph. 2019;71:30–39.
15. Zhang S, Liang G, Pan S, Zheng L. A fast medical image super resolution method based on deep learning network. IEEE Access. 2018;7:12319–27.
16. Bellver M, Maninis K-K, Pont-Tuset J, Giró-i-Nieto X, Torres J, Van Gool L. Detection-aided liver lesion segmentation using deep learning. 2017. arXiv preprint arXiv: 1711.11069.
17. Rashid Sheykhahmad F, Razmjooy N, Ramezani M. A novel method for skin lesion segmentation. International Journal of Information, Security Systems Management. 2015;4(2):458–66.
18. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. arXiv preprint arXiv: 1711.05225.
19. Liu H, Wang L, Nan Y, Jin F, Wang Q, Pu J. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. Comput Med Imaging Graphics. 2019;75:66–73.
20. Yao L, Prosky J, Poblenz E, Covington B, Lyman K. Weakly supervised medical diagnosis and localization from multiple resolutions. 2018. arXiv preprint arXiv: 1803.07703.
21. Wang H, Jia H, Lu L, Xia Y. Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. IEEE J Biomed Health Inform. 2019;24(2):475–85.
22. Guan Q, Huang Y. Multi-label chest X-ray image classification via category-wise residual attention learning. Pattern Recog Lett. 2020;130:259–66.
23. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: Transformers make strong encoders for medical image segmentation. 2021. arXiv preprint arXiv: 2102.04306.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Neural Information Processing Systems (NIPS). 2017; pp. 5998–6008.
25. Chen M, Radford A, Child R, Wu J, Jun H, Luan D, et al. Generative pretraining from pixels. Conference on Machine Learning (PMLR). 2020. pp. 1691–1703.
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020. arXiv preprint arXiv: 2010.11929.
27. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. 2020. arXiv preprint arXiv: 2010.04159.
28. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. 2021. arXiv preprint arXiv: 2103.14030.
29. Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jégou H, et al. LeViT: a vision transformer in convNet's clothing for faster inference. 2021. arXiv preprint arXiv: 2104.01136.
30. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. 2020. arXiv preprint arXiv: 2012.12877.
31. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. 2021. arXiv preprint arXiv: 2102.12122.
32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv: 1409.1556.
33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Conference on Computer Vision and Pattern Recognition (CVPR). 2016; pp. 770–778.
34. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Conference on Computer Vision and Pattern Recognition (CVPR). 2017; pp. 4700–4708.
35. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Conference on Computer Vision and Pattern Recognition (CVPR). 2015; pp. 1–9.
36. Wehrmann J, Cerri R, Barros R. Hierarchical multi-label classification networks. International Conference on Machine Learning (ICCV). 2018; pp. 5075–5084.
37. Freeman I, Roese-Koerner L, Kummert A. Effnet: an efficient structure for convolutional neural networks. 2018; arXiv preprint arXiv: 1801.06434.
38. Zhang M-L, Zhou Z-H, Engineering D. A review on multi-label learning algorithms. IEEE transactions on knowledge. 2013;26(8):1819–1837.
39. Durand T, Mehrasa N, Mori G. Learning a deep convnet for multi-label classification with partial labels. Conference on Computer Vision and Pattern Recognition (CVPR). 2019; pp. 647–657.
40. Krizhevsky, A, Sutskever, I, Hinton, G. ImageNet classification with deep convolutional neural networks. Proceedings of 26th Conference on Neural Information Processing Systems (NIPS). 2012; pp. 1097–1105.
41. Qin Z, Zhang P, Wu F, Li X. FcaNet: frequency channel attention networks. 2020; arXiv preprint arXiv: 2012.11879.
42. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Conference on Computer Vision and Pattern Recognition (CVPR). 2018; pp. 7132–7141.
43. Li X, Wang W, Hu X, Yang J. Selective kernel networks. Conference on Computer Vision and Pattern Recognition (CVPR). 2019. pp. 510–519.
44. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. 2020. arXiv preprint arXiv: 1910.03151.
45. Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional block attention module. IEEE International Conference on Computer Vision (ECCV). 2018. pp. 3–19.
46. Chen Y, Kalantidis Y, Li J, Yan S, Feng J. A^2-Nets: Double attention networks. 2018. arXiv preprint arXiv: 1810.11579.
47. Guha Roy A, Navab N, Wachinger C. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. In International conference on medical image computing and computer-assisted intervention. 2018; pp. 421–429.
48. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Conference on Medical image computing and computer-assisted intervention. 2015; pp. 234–241.
49. Elfwing S, Uchibe E, Doya K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw. 2018;107:3–11.
50. Hahnloser RH, Sarpeshkar R, Mahowald MA, Douglas RJ, Seung HS. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature. 2000;405(6789):947–51.
51. Islam MA, Jia S, Bruce ND. How much position information do convolutional neural networks encode?. 2020. arXiv preprint arXiv: 2001.08248.

52. NIH. ChestX-ray14 dataset. 2017. https://nihcc.app.box.com/v/ChestXray-NIHCC.
53. Kaggle. Catheter and Line Position Challenge. 2021. https://www.kaggle.com/c/ranzcr-clip-catheter-line-classification.
54. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vision. 2015;115(3):211–52.
55. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. 2016. arXiv preprint arXiv: 1608.03983.
56. Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017. arXiv preprint arXiv: 1711.05101.
57. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):1–8.