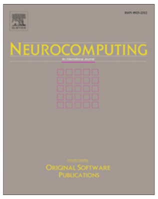
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



MANet: A two-stage deep learning method for classification of COVID-19 from Chest X-ray images

Yujia Xu, Hak-Keung Lam [*], Guangyu Jia

Centre for Robotics Research, Department of Engineering, King's College London, WC2R 2LS, UK

ARTICLE INFO

Article history:
Received 7 December 2020
Revised 5 February 2021
Accepted 10 March 2021
Available online 18 March 2021
Communicated by Zidong Wang

Keywords:
Chest X-ray images
COVID-19
Two-stage
Segmentation
Spatial attention
Convolutional neural networks

ABSTRACT

The early detection of infection is significant for the fight against the ongoing COVID-19 pandemic. Chest X-ray (CXR) imaging is an efficient screening technique via which lung infections can be detected. This paper aims to distinguish COVID-19 positive cases from the other four classes, including normal, tuberculosis (TB), bacterial pneumonia (BP), and viral pneumonia (VP), using CXR images. The existing COVID-19 classification researches have achieved some successes with deep learning techniques while sometimes lacking interpretability and generalization ability. Hence, we propose a two-stage classification method MANet to address these issues in computer-aided COVID-19 diagnosis. Particularly, a segmentation model predicts the masks for all CXR images to extract their lung regions at the first stage. A followed classification CNN at the second stage then classifies the segmented CXR images into five classes based only on the preserved lung regions. In this segment-based classification task, we propose the mask attention mechanism (MA) which uses the predicted masks at the first stage as spatial attention maps to adjust the features of the CNN at the second stage. The MA spatial attention maps for features calculate the percentage of masked pixels in their receptive fields, suppressing the feature values based on the overlapping rates between their receptive fields and the segmented lung regions. In evaluation, we segment out the lung regions of all CXR images through a UNet with ResNet backbone, and then perform classification on the segmented CXR images using four classic CNNs with or without MA, including ResNet34, ResNet50, VGG16, and Inceptionv3. The experimental results illustrate that the classification models with MA have higher classification accuracy, more stable training process, and better interpretability and generalization ability than those without MA. Among the evaluated classification models, ResNet50 with MA achieves the highest average test accuracy of 96.32% in three runs, and the highest one is 97.06%. Meanwhile, the attention heat maps visualized by Grad-CAM indicate that models with MA make more reliable predictions based on the pathological patterns in lung regions. This further presents the potential of MANet to provide clinicians with diagnosis assistance.

1. Introduction

CORONAVIRUS disease (COVID-19) has been declared as a pandemic by the World Health Organization (WHO) in March 2020 [52]. Till October 2020, the highly contagious COVID-19 has infected over 36 million people and caused more than one million deaths, and the numbers are still increasing at a rapid rate [14]. Hence, it is vital to diagnose the virus effectively to avoid its further spread.

In COVID-19 detection, three main screening methods are employed, including reverse transcriptase-polymerase chain reaction (RT-PCR) [51], computed tomography (CT) and chest X-ray (CXR). Compared with the other two methods, CXR imaging technique is more time-efficient, accessible and portable [51,2,16]. Furthermore, some researches have proved that CXR imaging can present radiographic abnormalities of COVID-19 positive cases [50,20]. These advantages make it an efficient imaging tool in the fight against the pandemic.

Since AlexNet [29] won ImageNet Large Scale Visual Recognition Challenge (ILSVRC) against other traditional methods by a large margin in 2012, deep learning (DL) techniques have increased rapidly. More advanced deep neural networks (DNNs) proposed later include VGG [41], GoogLeNet [44], ResNet [22], Inception [45], EfficientNet [46], etc. These state-of-the-art DNNs continuously improve their recognition ability in a variety of benchmark datasets by exploring the depth, width, resolution,

[*] Corresponding author.
E-mail addresses: yujia.xu@kcl.ac.uk (Y. Xu), hak-keung.lam@kcl.ac.uk (H.-K. Lam), guangyu.jia@kcl.ac.uk (G. Jia).

https://doi.org/10.1016/j.neucom.2021.03.034
0925-2312/© 2021 Elsevier B.V. All rights reserved.

structure of models. Meanwhile, DNNs have achieved impressive breakthroughs via incorporating some modules proposed in recent years like attention [54,17,48,15], multi-scale feature fusion [46,8], pseudo labeling [36], probabilistic generative modelling [56,18], biology-inspired designs [30,47], etc. These achievements have made DL techniques promising tools for developing real-world applications. Besides, the achievements in some domains have potential to be transferred to the other domains. For example, the infinite realistic data produced by generative models can be used for augmentation for improving classification accuracy [38,33,5], and self-supervised learning can help pre-train DL models achieve better performance in other tasks [21,10,9,19]. Closing the domain gaps can make domains beneficial from the development of the others, and jointly promote the DL researches. This motivates us to propose MANet to benefit classification from the trained segmentation models, based on the fact that classification depends only on the diagnosis-relevant features existing in the lung regions covered by segmentation masks.

In computer-aided diagnosis domain, an increasing number of DL applications have been proposed to assist the clinicians for a faster and more accurate diagnosis. This is because DL techniques like DNNs can learn the decision-making rules of given datasets efficiently, especially when sufficient data are available. Some researches have proved that DL has the power in assisting COVID-19 diagnosis. For example, [3] employed transfer learning on multiple convolutional neural networks (CNNs), including VGG19 [41], MobileNet [23], Inceptionv3 [45], Xception [11] and Inception-ResNetv2 [43], to distinguish COVID-19 positive cases from normal and bacterial pneumonia (BP), and achieved an accuracy of over 92% for all models. Another research in [1] modified the classic ResNet18 and claimed an accuracy of 95.12% in 3-class CXR classification (COVID-19, SARS and normal). Besides, some researches also presented promising results by combinations of DNNs and traditional machine learning techniques, e.g., decision tree [55] and SVM [40]. Furthermore, visual explanation tools like Grad-CAM++ [7] and layer-wise relevance propagation [4] (LRP) are applied in [27] to help localize the pathology of CXR images and making designed neural networks more convincing.

Although DL has achieved some success in computer-aided COVID-19 diagnosis, it still faces challenges, e.g., shortage of public CXR images, lack of interpretability and generalization ability [27,50,34]. Particularly, the performance of DL models depends heavily on the amount of training data. The data shortage caused by limited publicly available CXR data makes DL models likely to suffer from overfitting. Meanwhile, CXR images are often collected by different institutions with various radiographic devices and environments, and the public CXR datasets are specific for some certain diseases. Consequently, CXR images of different types differ in terms of data amount, radiographic features (e.g. illumination, contrast, resolution) and patient features (e.g. position, skeleton, age, etc.). The models trained on such an imbalanced dataset are likely to underfit the minor classes, especially for COVID-19 positive cases. Moreover, decisions made by the trained models may be based on the irrelevant radiographic and patient features as aforementioned, meaning that the attentions of models are not in pathological regions, or even not in lung regions. This makes the models lack of interpretability, and hard to be generalized to new or unusual samples. These mentioned problems inspire us to use CNN-based segmentation methods to extract main lung regions to minimize the differences between CXR images from different repositories. A following classification model can then enhance their generalization ability by focusing their attention only on the extracted lung regions with less diagnosis-irrelevant features.

To develop solutions for the pandemic and address the mentioned issues, this paper introduces MANet, a two-stage DL method, to provide clinicians with efficient lung disease classification and basis of computer-aided determination. The basis indicating the potential pathological regions can assist clinicians in localization and detection of pathological changes, and can be visualized by tools like Grad-CAM [39]. MANet works for five-class classification, including normal, COVID-19, tuberculosis (TB), bacterial pneumonia (BP), and viral pneumonia (VP), respectively. Specifically, at the first stage in MANet, the segmentation model takes the original CXR images as inputs and predicts the corresponding lung masks. A followed CNN with mask attention mechanism (MA) classifies the segmented CXR images into five classes based only on the preserved lung regions. MA proposed in this paper is a kind of undifferentiable soft spatial attention [53,32] mechanism in CNNs that uses the predicted masks from the first stage as spatial attention maps to adjust the features in CNNs at the second stage. The spatial attention map in MA for a segmented CXR image is initialized as the corresponding predicted lung mask and extends via iterative average pooling with specific parameters. We use MA to adjust the spatial attention of classification models and stabilize their training process. The two stages of MANet are necessary as they play different roles in this segment-based classification task. The segmentation model predicts the masks of CXR images via which only the lung regions are preserved and the diagnosis-irrelevant features mentioned in the previous paragraph outside the lung regions are filtered out. The classification model at the second stage employs MA to concentrate its attention on the preserved informative lung regions.

In the experiments, we select UNet [37] with ResNet backbone (ResUNet) as the segmentation model to segment out the lung regions of all CXR images, and perform classification over the segmented CXR images using four classic CNNs with or without proposed MA, including ResNet34, ResNet50, VGG, and Inceptionv3. Besides, we also evaluate the results of classification models with CBAM [54], a commonly used soft attention module, to emphasize the advantage of MA in this segment-based classification task. To fairly compare all the models, we train and evaluate them over the same test settings in three trails, with unique random seeds. Eventually, ResNet50 with MA as the classification model achieves the highest average test accuracy of 96.32% and the highest test accuracy in three trails is 97.06%. In addition, the experimental results demonstrate that the classification models with MA surpass both models without MA and models with CBAM in most cases in terms of test accuracy. Furthermore, we employ Grad-CAM [39] to visualize the attention of all involved classification models. The attention heat maps illustrate that the attentions of classification models with MA are in pathological regions while those of others are more disordered, indicating MANet has better interpretability.

The contributions of this work are listed as follows:

a) MANet proposes a way to benefit classification from appropriate segmentation, as multiple experiments demonstrate that the usage of MA improves classification accuracy.
b) MANet shows improved interpretability and generalization ability as its predictions are based on the segmented lung regions of CXR images, while the diagnosis-irrelevant regions are filtered out after the first stage.
c) MA can stabilize the training of CNNs in this segment-based classification task. It reduces the fluctuations of both loss and accuracy, and trains models more stably.
d) MA is a light attention module that can be easily incorporated into most CNNs, both adjusting the extracted features and localizing attention of CNNs in desired regions. The four CNNs, including ResNet34, ResNet50, VGG16, and Inception-v3, improve their test classification accuracy by applying MA with 5.78% to 21.55% running time increase.

The rest of the paper is organized as follows: Section 2 presents the collection of CXR datasets, pre-processing, and post-processing. Section 3 introduces the entire workflow of MANet and the proposed spatial attention module MA. The corresponding experimental results and analysis are illustrated in Section 4, as well as the visualization of attention heat maps. Section 5 summarizes the work and discusses the future directions.

2. Dataset collection and processing

2.1. Public data repositories

To improve the classification accuracy and generalization ability of DL models for COVID-19 diagnosis, we collect CXR images as much as possible and apply data augmentation to increase the data diversity. In this research, the employed dataset is a combination of three public CXR data repositories, including a combined CXR dataset contributed by both Montgomery County and Shenzhen No. 3 People's Hospital [26], a CXR dataset released by Kermany et al. [28], and a public open dataset on GitHub specific for COVID-19 [13,12]. The combined dataset contains CXR images in five classes, normal, COVID-19, TB, BP and VP with 1840, 433, 394, 2780 and 1345 images, respectively.

As can be seen in Fig. 2, the two models in MANet require different data processing techniques. Specially, the segmentation model is trained over pairs of CXR images and the corresponding lung masks. And the pairs of segmented CXR images and the corresponding post-processed masks are the inputs for the classification model. The inputs for both two stages require pre-processing, and the outputs at the first stage require to be post-processed to remove some prediction defects before going to the next stage. In the next two subSections 2.2 and 2.3, we demonstrate the details of the dataset composition and image processing at both two stages. It is worth noting that we explain the dataset-relevant contents thoroughly in this section so that Section 3 can be focused on technical details.

2.2. Dataset for segmentation and data processing

To the best of the authors' knowledge, no public datasets include all five classes of CXR images. Hence, images in different classes are often from different datasets and have different characteristics. As presented in Fig. 1, images from multiple datasets differ in terms of some features, e.g. patient position, skeletons, image resolution, radiographic illumination, etc. DL models trained on such a dataset may classify images according to these diagnosis-irrelevant features, rather than the informative pathological patterns.

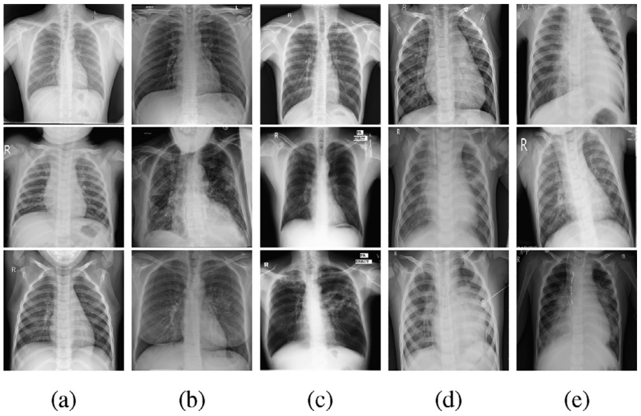


Fig. 1. Example CXR images in five classes (Each column contains three examples in one class). (a): Normal, (b): COVID-19, (c): TB, (d): BP, (e): VP.

To alleviate the irrelevant biases among these data repositories, we segment out the lung regions of all CXR images at the first stage. At this stage, we trained a classic segmentation model UNet with residual connections [22] (ResUNet) to segment out the lung regions of all images automatically. Since the pixel-wise labelling for image masks takes much time and efforts from specialists, only 359, 345 and 202 pairs of images and masks for classes normal, TB and COVID-19 are collected from the datasets [26,13,12] while image-mask pairs for the other two classes are unavailable. Among these data, 323, 307 and 185 image-mask pairs for classes normal, TB, and COVID-19 are included for training the ResUNet. And the rest 10% are used as the test set. Besides, the training process also involves data pre-processing including resizing, normalization, and augmentation. Specifically, all CXR images are resized to 512 × 512 resolution and normalized to a range $[0,1]$ according to min–max feature scaling [25]. And the on-the-fly augmentation implemented by Albumentations library [6] are listed as follows:

1. Shifting: shift the images horizontally and vertically randomly with the ratio range $[-6.25\%, 6.25\%]$.
2. Scaling: scale the images randomly with the ratio range $[-20\%, 20\%]$.
3. Rotation: rotate the images by angles selected randomly from the uniform distribution in $[-10, 10]$.

All these on-the-fly augmentations with specific parameters are determined by numerous experiments and have a probability of 50% of being applied or not for training samples. The evaluation step does not involve augmentation.

Post-processing for masks at the end of the first stage is required as the ResUNet segmentation is not always perfect. In some cases, the predicted masks for the left and right lungs are concatenated and there might be small defects outside the lung regions. Hence, we applied two post-processing operations implemented by Scikit-image [49], filling holes and removing small objects, to fine-tune the predicted masks. As demonstrated in Fig. 3, the post-processed masks for all classes contain less detects. These posterior operations provide better training inputs for the classification models at the second stage.

2.3. Dataset for classification and data processing

Once the segmentation and following post-processing at the first stage are done, we can then perform classification over the pairs of post-processed segmented images and masks as demonstrated in the last row in Fig. 3. It should be noted that we split them into three sets, training, validation and test, according to the ratio $8:1:1$. The numbers of samples in all classes for classification are summarized in Table 1.

To train the classification models with or without MA, the pairs of post-processed segmented images and masks require pre-processing as well for normalization and diversity enhancement, similar to that for segmentation. Specifically, we implement the data augmentation operations via Albumentations library [6]. The operations and corresponding ratio ranges determined by numerous experiments are listed as follows:

1. Shifting: shift the pairs of images and masks horizontally and vertically randomly with the ratio range $[-6.25\%, 6.25\%]$.
2. Scaling: scale the pairs of images and masks randomly with the ratio range $[-20\%, 20\%]$.
3. Rotation: rotate the pairs of images and masks by angles selected randomly from the uniform distribution in $[-30, 30]$.

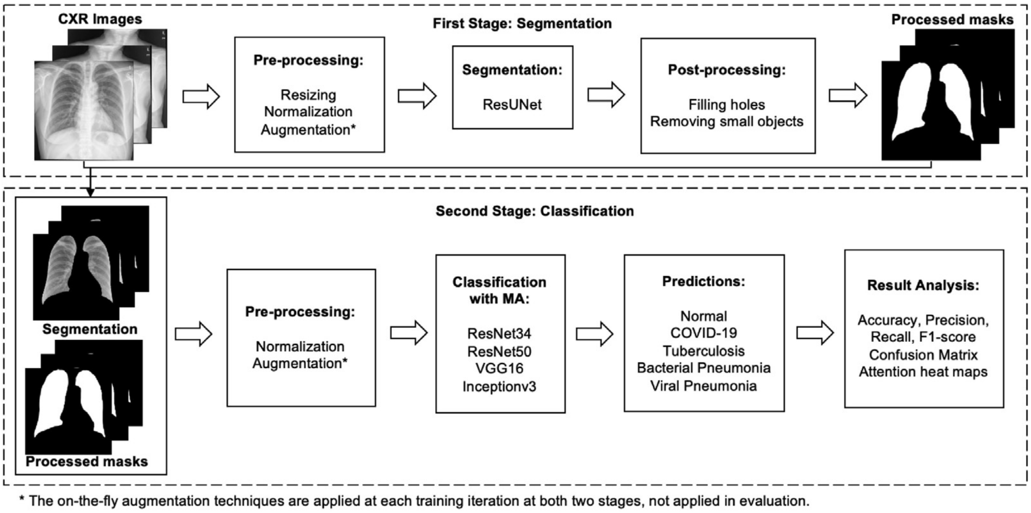


Fig. 2. Schematic representation of the training workflow of MANet.

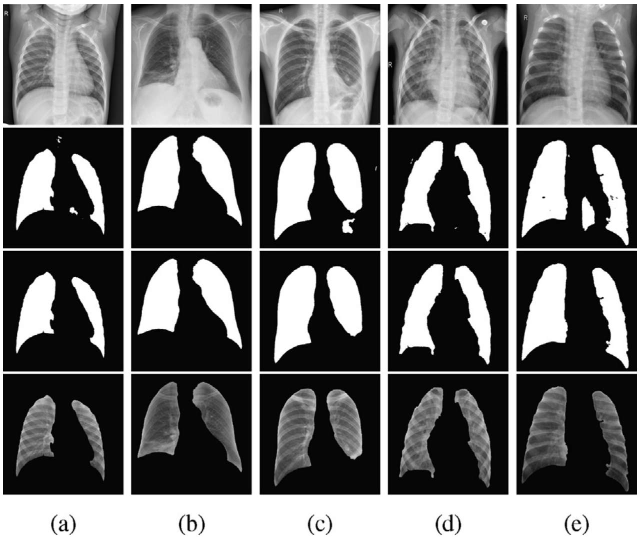


Fig. 3. Example processed images in class: (a) Normal, (b): COVID-19, (c): TB, (d): BP, (e): VP. The images in four rows represent original images, predicted masks, post-processed masks, and final segmentation.

4. Adjustion of brightness and contrast: adjust the brightness and contrast of images with the ratio range $[-20\%, 20\%]$.

These on-the-fly augmentations have a probability of 50% of being applied to the training pairs of images and masks, both for diversity enhancement and avoiding over-augmentation. Besides, the augmented images are normalized to a common scale $[-1, 1]$ according to min–max feature scaling [25].

Table 1
Distribution of samples for classification in all infection types.

Set	Normal	COVID-19	TB	BP	VP
Training	1484	331	306	2239	1073
Validation	185	51	50	258	135
Test	171	51	38	283	137
Total	1840	433	394	2780	1345

3. Two-stage segment-based classification

3.1. Workflow of MANet

In this work, we introduce a segment-based classification method named MANet. Its integrated workflow is illustrated in Fig. 2. Unlike the end-to-end classification CNNs, MANet contains two stages that are segmentation and classification. Generally, the segmentation at the first stage is to segment out the diagnosis-relevant lung regions that the classification models at the second stage need to concentrate on for predicting the classes of inputs.

Attention mechanisms, especially soft attention, are being used more and more frequently since many works [54,24,17] have revealed that attention can expand capabilities of networks and allow approximating more complicated functions. However, the conventional differentiable soft attention is computationally expensive and often over-parameterized. MA proposed in this work reduces the computational cost by defining the attention maps for all features prior to feature extraction in CNNs via a segmentation model trained at the first stage, and also distributes the attentions of CNNs in the segmented lung regions. The architectures for the involved CNNs with MA and corresponding hyperparameters are demonstrated in the following two subsections.

3.2. Stage 1: Segmentation

At the first stage, we employ the classic segmentation model UNet [37] with ResNet backbone (ResUNet) to segment out the lung regions of all CXR images. Its architecture and the residual connection in all convolutional blocks are demonstrated in Figs. 4a and 4b, respectively. The symmetric architecture of ResUNet can extract features in different levels from low to high, from which the corresponding lung masks can be predicted more precisely. Meanwhile, the residual connection avoids ResUNet from going too deep and also alleviates the gradient diminishing problem.

3.3. Stage 2: Classification with MA

In nearly all CNN-based image classification tasks, the predictions are based on the features extracted from the entire images. Such a decision-making process is likely to be affected by the diagnosis-irrelevant features of the given inputs. These features,

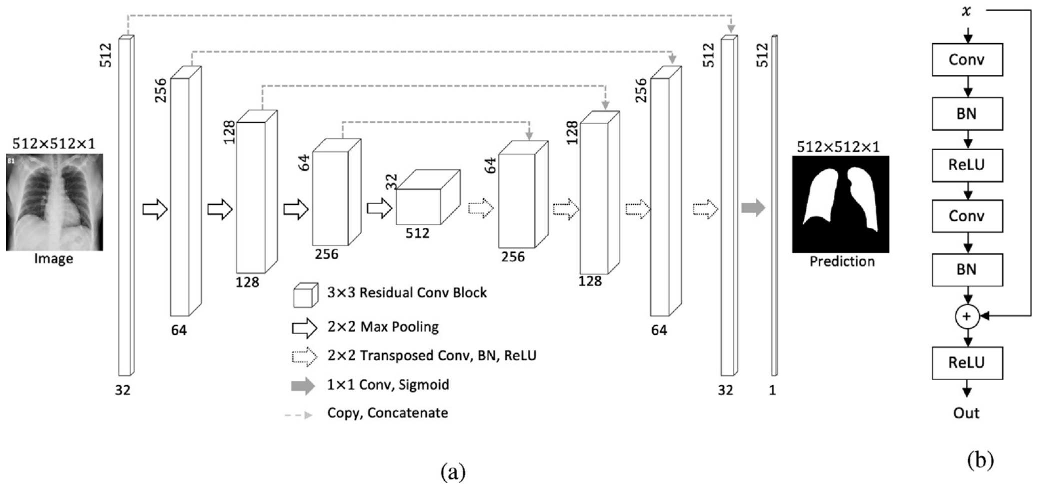


Fig. 4. Architecture of (a): ResUNet, (b): a residual convolution block in ResUNet.

e.g., areas of dark regions, edges of lungs, etc., may mislead the CNNs and cause unstable training. Specifically, in medical disease classification, CNNs only need to concentrate on local pathological regions (lung regions in CXR classification). Hence, the masks predicted by the segmentation model in Section 3.2 can be employed as the pre-defined regions for CNNs to focus on. We propose MA to guide the CNNs at the second stage to concentrate on the lung regions covered by the predicted masks.

Most CNNs are implemented by stacking convolutional blocks composed of convolution, normalization and activation layers. In these models, the proposed MA can be conveniently applied by adjusting the feature values in each block according to calculated attention maps. Fig. 6 demonstrates a classic convolutional block with MA. Essentially, MA is to adjust the feature values in CNNs based on a spatial attention map indicating the amount of valid information covered in their receptive fields. It forces the spatial attention of CNNs located in the desired regions unchanged while suppressing that in other regions.

The overall transformation of features in a convolutional block with MA can be written as Eq. (1). Mathematically, in a CNN with MA, an input is a pair of segmented CXR image I_0 and corresponding mask M_0 predicted by the segmentation model. Likewise, in its i^{th} convolutional block, the input is a pair of 3-D input feature matrix I_i, and its corresponding 2-D spatial attention map M_i. M_i can be seen as the spatial attention map for I_i, and an element $M_i(x,y) \in [0,1]$ in the map indicates the amount of available information in $I_i(x,y)$. An unbiased convolutional layer, with kernel size = k, stride = s and padding = p, then transforms I_i to $f(I_i;k,s,p)$. The following BN layer normalizes the convolutional value to $\hat{f}(I_i;k,s,p)$. When the receptive fields of extracted features changes in some operations (e.g. pooling, convolution with kernel size > 1, etc.), the spatial attention map evolves jointly via a consistent average pooling, $M_{i+1} = P(M_i;k,s,p)$. MA in a convolutional block adjusts the normalized features based on the corresponding spatial attention map through an element-wise multiplication denoted by $\odot$, after which a non-linear activation function $\mathscr{A}$ arises.

$$\begin{aligned} I_{i+1} &= \mathscr{A}(\hat{f}(I_i;k,s,p) \odot P(M_i;k,s,p)) \\ M_{i+1} &= P(M_i;k,s,p) \end{aligned} \tag{1}$$

The proposed MA can be easily applied in most CNNs as illustrated in Fig. 5. Specifically, for a common convolutional block like in VGG, MA is implemented by multiplying the normalized features with calculated attention maps. While in the ResNet family, MA might be affected by the skip-connection operations. This negative impact is slight as the difference between the two attention maps in two connected blocks is trivial. The deep and residual features in ResNet basic blocks are added up, producing fused features from two different receptive fields. The fused features cause uncertainty when choosing receptive field for calculating attention maps for MA. We choose to generate spatial mask attention maps according to the larger receptive field from deep features and the classification results in Section 4 can prove that MA works for ResNet in this way. The bottleneck structure in deep ResNet only performs MA once because 1×1 convolution does not change the receptive field of extracted features. The inception module is different from the preceding ones as features in different channels have different receptive fields. In this case, the spatial attention map in MA is simply calculated based on the largest receptive field of features in all channels.

Essentially, MA is a kind of spatial attention mechanism but different from others in terms of its nondifferentiability and independence of target features. In other soft attention mechanisms, like CBAM [54] illustrated in Fig. 7, the attention is inferred from the feature values and require trainable parameters, while that in convolutional blocks with MA is independent on the feature values. MA decouples the spatial attention generation from feature extraction and avoids training attention modules repeatedly. Spatial attention in MA is to suppress only the outlier-features whose receptive fields are outside the lung regions. Since the lung regions are already predicted at the first stage, the attention maps in MA do not need to be re-calculated according to extracted features at the second stage. The attention maps evolve progressively based on the receptive fields of features rather than their values.

4. Results and analysis

In this section, we demonstrate the experimental settings, the results of both two stages in MANet, and compare the performance of models with or without MA. Besides, Grad-CAM [39] is employed to help visualize the attentions of involved CNNs.

4.1. Results for segmentation

At the first stage, we trained the ResUNet on the dataset (in Section 2.2) using the Adam optimizer for 50 epochs to minimize the dice loss [42]. The hyper-parameters are as follows: batch size = 64,

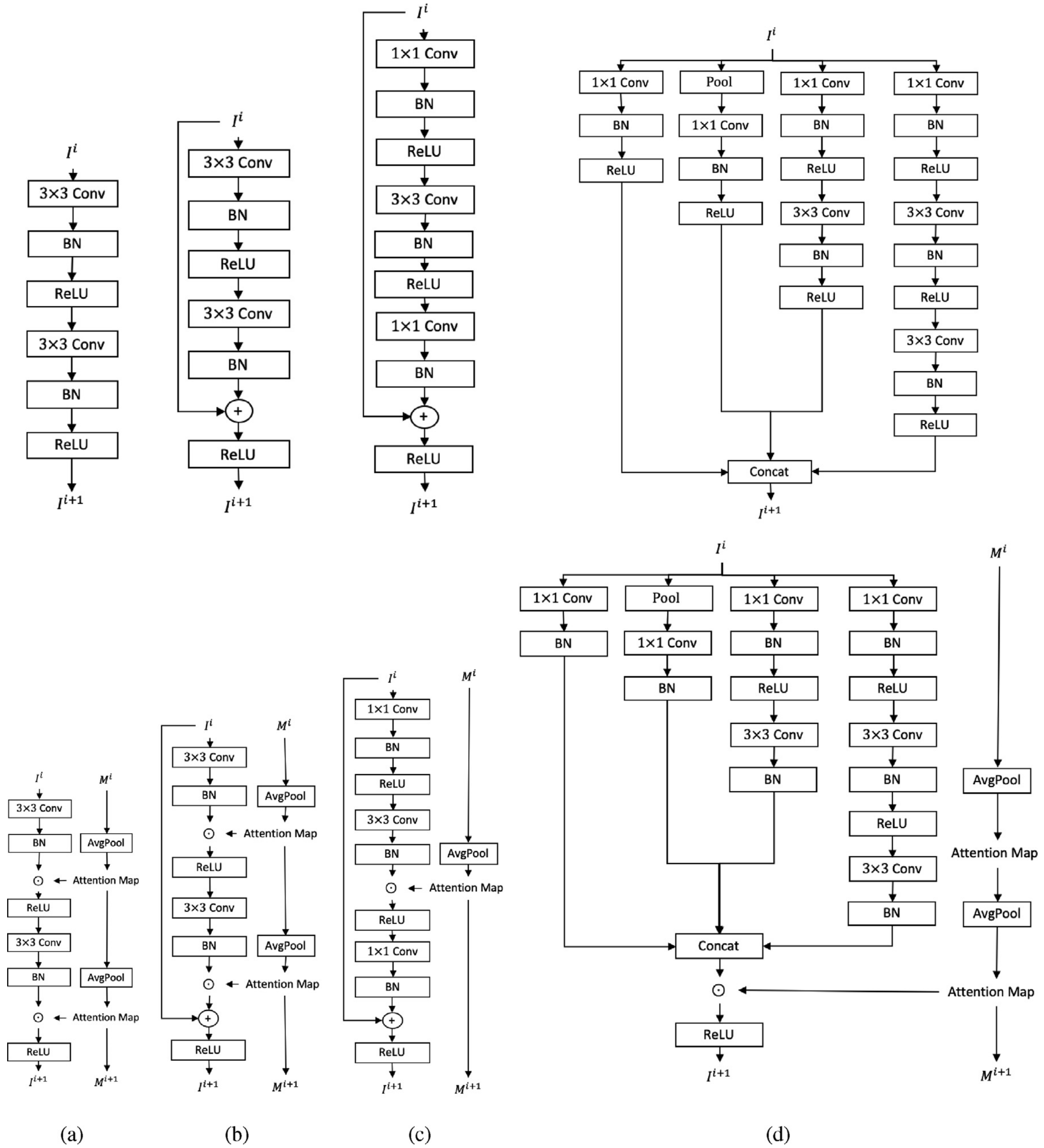


Fig. 5. Structures of (a): convolutional block in VGG, (b): basic convolutional block in ResNet, (c): bottleneck in ResNet, (d): Inception module in Inceptionv3. (Top: original convolutional blocks without MA. Bottom: convolutional blocks with MA.).

constant learning rate = 2×10^{-4}, weight decay = 1×10^{-5}. Besides, the trainable weights of the involved convolution layers are initialized according to Kaiming Normal [22] while the weights of batch normalization (BN) layers are initialized to one with zero bias.

As illustrated in Section 2.2, only 359, 345 and 202 CXR image-mask pairs for classes normal, TB and COVID-19 are available. Hence, at the first stage of MANet, we train and evaluate the ResUNet segmentation model over the dataset containing these

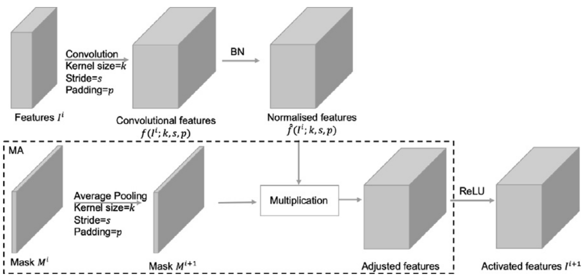


Fig. 6. Schematic representation of a classic convolutional block with MA.

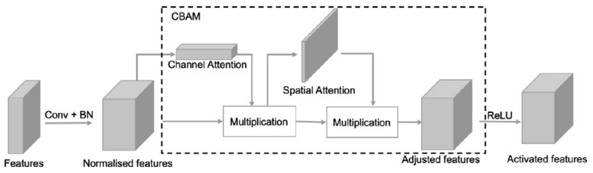


Fig. 7. Schematic representation of a convolutional block with CBAM.

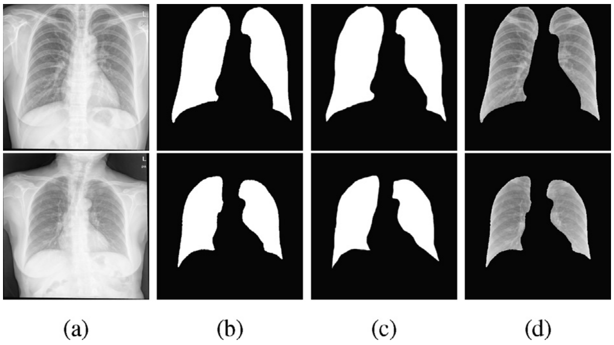


Fig. 8. Visualization results of ResUNet. The images from left to right represent the original CXR images, masks predicted by trained ResUNet, ground-truth masks, and segmentation, for a (top) normal and a (bottom) TB cases in the test set.

Table 2
Test accuracy of original models, models with MA, and models with CBAM.

Model	Trail 1	Trail 2	Trail 3	Average
ResNet34	**96.76%**	95.44%	95.15%	95.78%
ResNet34(MA)	96.18%	**96.32%**	95.59%	**96.03%**
ResNet34(CBAM)	95.29%	93.82%	**96.18%**	94.92%
ResNet50	95.59%	**96.47%**	96.18%	96.08%
ResNet50(MA)	**96.47%**	95.44%	**97.06%**	**96.32%**
ResNet50(CBAM)	96.18%	95.74%	94.85%	95.59%
VGG16	94.56%	93.37%	93.68%	93.87%
VGG16(MA)	**95.00%**	**94.56%**	**95.74%**	**95.10%**
Inceptionv3	94.41%	95.44%	**96.32%**	95.39%
Inceptionv3(MA)	**96.62%**	**96.03%**	95.44%	**96.03%**

The text in bold indicates that the corresponding model achieves the highest test accuracy among all models with one backbone (ResNet34, ResNet50, VGG16 or Inceptionv3) in one trail.

Table 3
Comparison of the involved models in terms of computational space and time.

Model	Params (M)	FLOPs (M)	Running time	Time increase (%)
ResNet34	21.28	18766.76	1h56m	–
ResNet34(MA)	21.28	18766.99	2h21m	21.55
ResNet34(CBAM)	21.44	18781.39	2h43m	40.52
ResNet50	23.51	21058.83	2h37m	–
ResNet50(MA)	23.51	21058.98	2h47m	6.37
ResNet50(CBAM)	26.03	21099.81	3h55m	49.68
VGG16	134.29	80208.35	7h30m	–
VGG16(MA)	134.29	80209.15	7h56m	5.78
Inceptionv3	21.80	17537.14	2h41m	–
Inceptionv3(MA)	21.80	17537.40	2h56m	9.32

Table 4
Classification results for MANet using ResNet50 with MA as the classification model.

Class	Precision	Recall	F1-score	Support
Normal	0.94	0.96	0.95	171
COVID-19	0.94	0.96	0.95	51
TB	0.91	0.79	0.85	38
BP	1.00	0.99	0.99	283
VP	0.98	0.99	0.99	137
Weighted Average	0.97	0.97	0.97	680

Table 5
Confusion matrix for ResNet50 with MA.

		Predicted				
		Normal	COVID	TB	BP	VP
Actual	Normal	165	0	3	0	3
	COVID	1	49	0	1	0
	TB	8	0	30	0	0
	BP	0	3	0	280	0
	VP	1	0	0	0	136

image-mask pairs. From the segmentation results presented in Fig. 8, the predicted masks are similar with their ground truth, indicating the trained ResUNet can correctly segment out the lung regions.

Regarding objective evaluation for ResUNet segmentation, the intersections over union (IoUs) for test samples in three classes (36 normal, 38 TB and 17 COVID-19) are measured. We exclude the other two classes, BP and VP, at the first stage, because their corresponding image-mask pairs are unavailable in the collected data repositories. The conditional average test IoUs for classes normal, TB and COVID-19 are $93.49\%, 93.14\%$ and 89.07%, respectively. And the overall average test IoU is 92.50%. Before the next stage, we predicted the masks for all obtained CXR images followed by the post-processing illustrated in Section 3.2. The examples of the post-processed images are demonstrated in the last row in Fig. 3.

4.2. Results for classification

In our experiments, we trained the involved models, including the original models, models with MA, and models with CBAM, with the same setting of hyper-parameters and training policy. The employed optimizer is SGD with an initial learning rate of 0.01 decayed to 0 according to a cosine annealing scheduler [31], and a momentum$=0.9$. It aims to minimize the cross entropy loss over 200 epochs. Besides, we implemented the distributed data parallelism of all involved models on four NVIDIA Tesla P100 GPUs via Pytorch [35] framework to accelerate and improve training. With the parallel models, the batch size is enlarged to 64 to improve the classification results.

At the second stage, we trained the four types of models with or without MA as the classification model, including VGG16, ResNet34, ResNet50 and Inceptionv3, over the dataset after ResUNet segmentation and post-processing. The composition and representative samples of the pairs of post-processed segmented images and masks at this stage are shown in Section 2.3. To evaluate the performance of models, we trained and tested all models on the same dataset in three trails.

Table 2 presents the test accuracy for all tested models in three training trails. All models are trained with the post-processed CXR images as inputs. From this summary table, all models with MA surpass the original ones in terms of the average test accuracy, and MA improves the test accuracy of the original models in 9 out of 12 experiments. Among these models, ResNet50 with MA as the classification model achieves the best average test accuracy 96.32%. Meanwhile, the best test accuracy obtained by ResNet50 with MA is 97.06% in the third trail. Besides, CBAM is also employed in ResNet34 and ResNet50 for comparison with MA. The results show that models with CBAM cannot even outperform the original models without MA in terms of the average test accuracy.

MA is a light attention module that costs no extra parameters and can be easily applied in most mainstream CNNs. We present the numbers of trainable parameters, floating point operations

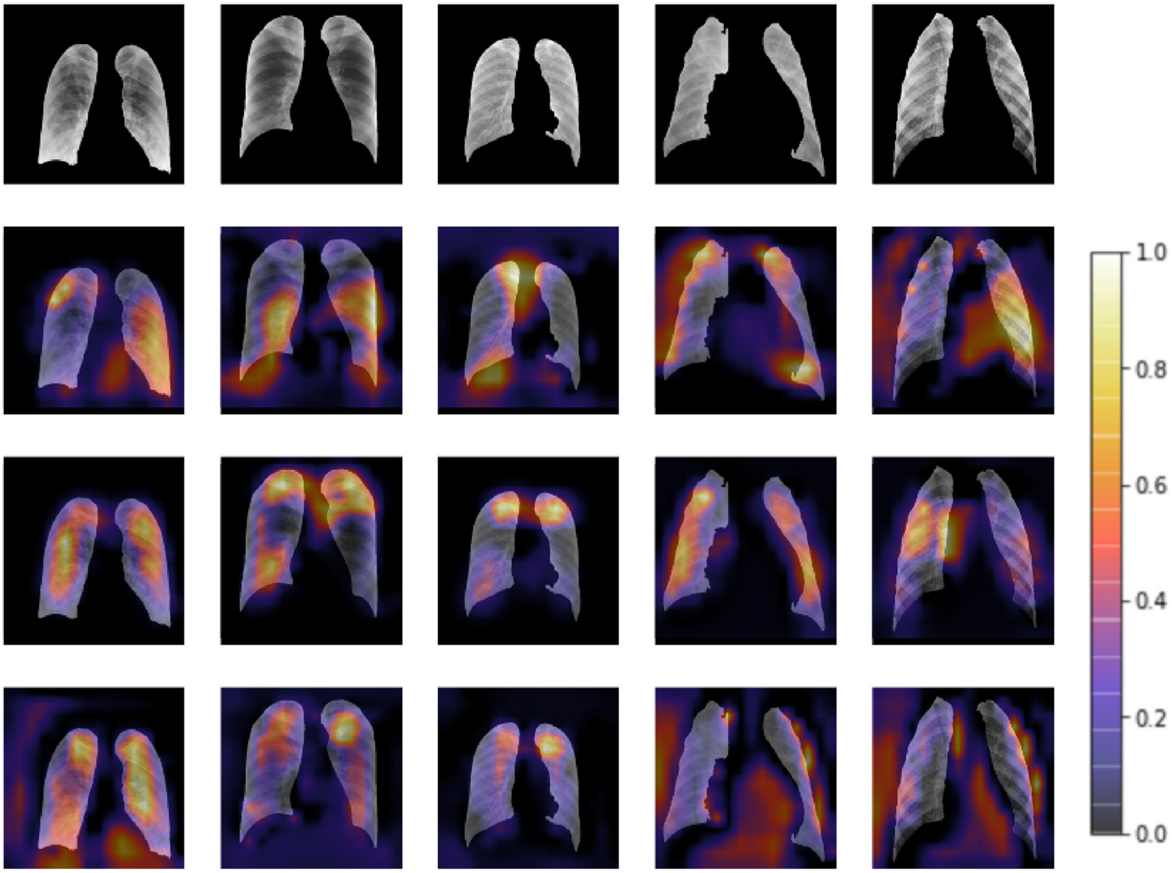


Fig. 9. Segmented CXR images and the corresponding attention heat maps visualized by Grad-CAM. The images in the first row are in five classes: COVID-19, normal, TB, BP and VP, respectively. The later three rows of attention heat maps are for original ResNet50, ResNet50 with MA and ResNet50 with CBAM, respectively. (Only the attentions for the ground truth classes are highlighted.) The color bar on the right hand indicates the attention values.

(FLOPs), and running time of all involved models in Table 3. It is worth noting that MA models achieve the best average test accuracy with trivial extra FLOPs and with 5.78% to 21.55% extra running time.

Table 4 presents the classification report for the best model, ResNet50 with MA in the third trail, with the corresponding confusion matrix shown in Table 5. It can be drawn from the statistics that the classification performance is proportional to the amount of data since that for dominated classes (BP and VP), outperforms others.

Besides, we also plot the training process of all models in one trail as shown in Fig. 10. It can be noted that MA can stabilize the training as the sharp ups and downs of validation accuracy are significantly reduced with the usage of MA, especially in the later phases. Furthermore, the validation accuracy of models with MA improves more quickly than that of models with CBAM, indicating that the decoupling of spatial attention generation and feature extraction by MA can accelerate the convergence of CNNs.

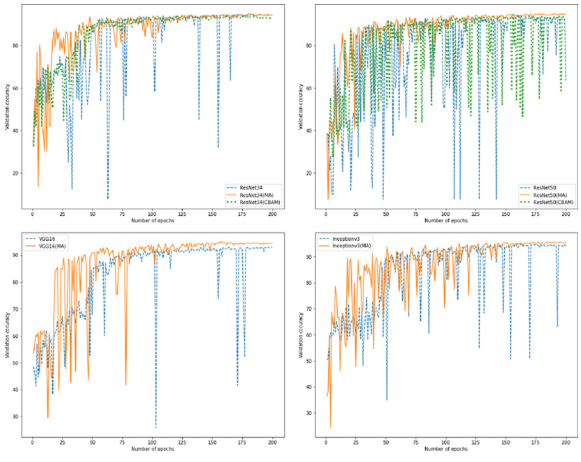


Fig. 10. Validation accuracy of all models over 200 epochs.

4.3. Attention visualization

MA is a light module to concentrate the attentions of CNNs on the regions pre-defined by the segmentation model. To show its availability, we plot the attention heat maps visualized by Grad-CAM [39] in Fig. 9, presenting the attentions in the last bottleneck blocks of all trained models extended from ResNet50. It can be noted from the second row in Fig. 9 that the attentions of the original model are mainly distributed around the segmentation edges. After applying CBAM, the distribution of attentions becomes more scattered and irregular. Comparing with others, the models with MA present relatively more practical attentions located in the segmented lung regions. It indicates that MANet has a better generalization ability and interpretability as MANet classifies the CXR samples based on the diagnosis-relevant features. The attention heat maps improved by the application of MA show that MANet has the potential to assist COVID-19 diagnosis.

5. Conclusion

This paper proposes the MANet, a new two-stage classification method for the classification of COVID-19 positive cases from CXR images. The dataset used in this study is collected from three data repositories, containing 6792 CXR images in five classes including normal, COVID-19 positive, TB, BP and VP (expect for COVID-19). MANet involves two stages that are segmentation and classification, introducing a new spatial attention module MA to concentrate the attention of classification model on the regions predicted by the segmentation model. We implement the segmentation through a UNet model with ResNet backbone, and apply MA into four classic CNNs for classification, including ResNet34, ResNet50, VGG16 and Inceptionv3. The statistical results and training demonstrate that the spatial attention mechanism MA in MANet can improve both the classification performance and training stability of the original models. Moreover, when comparing the attention heat maps of models with or without MA, the heat maps of models with MA indicate the potential pathological regions of CXR images which enhances the interpretability of MANet. Further directions include developing end-to-end MANet to improve its efficiency, enhancing the robustness by collecting more COVID-19 CXR images, and extending this work to other segment-based classification tasks.

CRediT authorship contribution statement

Yujia Xu: Conceptualization, Methodology, Software, Writing - original draft. **Hak-Keung Lam:** Supervision, Writing - review & editing. **Guangyu Jia:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank King's College London for their support and University of Cambridge Research Computing service funded by EPSRC Tier-2 capital grant EP/P020259/1 for provided computational resources.

References

[1] A. Abbas, M.M. Abdelsamea, M.M. Gaber, Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network, 2020, ArXiv Preprint ArXiv:2003.13815..

[2] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology (2020) 200642..

[3] I.D. Apostolopoulos, T.A. Mpesiana, COVID-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks, Physical and Engineering Sciences in Medicine 1 (2020).

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS One 10 (2015).

[5] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D.A. Dickie, M. V. Hernández, J. Wardlaw, D. Rueckert, GAN augmentation: Augmenting training data using generative adversarial networks, 2018, ArXiv Preprint ArXiv:1810.10863..

[6] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A.A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2020), https://doi.org/10.3390/info11020125.

[7] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 839–847.

[8] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2017) 834–848.

[9] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[10] X. Chen, K. He, Exploring simple siamese representation learning, 2020, ArXiv Preprint ArXiv:2011.10566..

[11] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[12] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, 2020, ArXiv: 2003.11597 https://github.com/ieee8023/covid-chestxray-dataset..

[13] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, COVID-19 image data collection: Prospective predictions are the future, 2020, ArXiv: 2006.11988 https://github.com/ieee8023/covid-chestxray-dataset..

[14] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, The Lancet Infectious Diseases 20 (2020) 533–534.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, ArXiv Preprint ArXiv:2010.11929..

[16] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest CT for COVID-19: comparison to RT-PCR, Radiology (2020) 200432..

[17] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[18] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, ArXiv Preprint ArXiv:1406.2661..

[19] J.B. Grill, F. Strub, F. Altché, C. Tallec, P.H. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z.D. Guo, M.G. Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, 2020, ArXiv Preprint ArXiv:2006.07733.

[20] W. Guan, Z. Ni, Y. Hu, W. Liang, C. Ou, J. He, L. Liu, H. Shan, C. Lei, D.S. Hui, et al., Clinical characteristics of coronavirus disease 2019 in China, New England Journal of Medicine 382 (2020) 1708–1720.

[21] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[23] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, ArXiv Preprint ArXiv:1704.04861.

[24] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 mb model size, 2016, ArXiv Preprint ArXiv:1602.07360.

[25] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, ArXiv Preprint ArXiv:1502.03167.

[26] S. Jaeger, S. Candemir, S. Antani, Y.X.J. Wáng, P.X. Lu, G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, Quantitative Imaging in Medicine and Surgery 4 (2014) 475.

[27] M. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan, et al., Deepcovidexplainer: Explainable COVID-19 predictions based on chest X-ray images, 2020, ArXiv Preprint ArXiv:2004.04582.

[28] D. Kermany, K. Zhang, M. Goldbaum, Labeled optical coherence tomography (OCT) and chest X-ray images for classification, Mendeley Data 2 (2018).

[29] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105.

[30] H. Liu, L. Ma, Z. Wang, Y. Liu, F.E. Alsaadi, An overview of stability analysis and state estimation for memristive neural networks, Neurocomputing 391 (2020) 1–12.

[31] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, 2016, ArXiv Preprint ArXiv:1608.03983.

[32] M.T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, 2015, ArXiv Preprint ArXiv:1508.04025.

[33] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, C. Malossi, Bagan: Data augmentation with balancing GAN, 2018, ArXiv Preprint ArXiv:1803.09655.

[34] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, G.J. Soufi, Deep-covid: Predicting COVID-19 from chest X-ray images using deep transfer learning, 2020, ArXiv Preprint ArXiv:2004.09363.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8026–8037.

[36] H. Pham, Q. Xie, Z. Dai, Q.V. Le, Meta pseudo labels, 2020, ArXiv Preprint ArXiv:2003.10580.

[37] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[38] V. Sandfort, K. Yan, P.J. Pickhardt, R.M. Summers, Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks, Scientific Reports 9 (2019) 1–9.

[39] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization,

in: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.
[40] P.K. Sethy, S.K. Behera, Detection of coronavirus disease (COVID-19) based on deep features, 2020, p. 2020, Preprints 2020030300.
[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, ArXiv Preprint ArXiv:1409.1556.
[42] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M.J. Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 240–248.
[43] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, 2016, ArXiv Preprint ArXiv:1602.07261.
[44] 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 Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[46] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[47] A. Thomas, Memristor-based neural networks, Journal of Physics D: Applied Physics 46 (2013) 093001.
[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, ArXiv Preprint ArXiv:1706.03762.
[49] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, the scikit-image contributors scikit-image: image processing in Python, PeerJ 2 (2014), https://doi.org/10.7717/peerj.453, DOI: 10.7717/peerj.453 e453.
[50] L. Wang, A. Wong, COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images, 2020, ArXiv Preprint ArXiv:2003.09871.
[51] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, W. Tan, Detection of SARS-CoV-2 in different types of clinical specimens, Jama 323 (2020) 1843–1844.
[52] WHO, WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19 – 11-march-2020..
[53] Wikipedia contributors, Visual spatial attention – Wikipedia, the free encyclopedia, 2020. https://en.wikipedia.org/w/index.php?title=Visual_spatial_attention&oldid=975937421. [Online; accessed 24-November-2020]..
[54] S. Woo, J. Park, J.Y. Lee, I. So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[55] S.H. Yoo, H. Geng, T.L. Chiu, S.K. Yu, D.C. Cho, J. Heo, M.S. Choi, I.H. Choi, C. Cung Van, N.V. Nhung, et al., Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging, Frontiers in Medicine 7 (2020) 427.
[56] N. Zeng, Z. Wang, H. Zhang, W. Liu, F.E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, Cognitive Computation 8 (2016) 684–692.

Yujia Xu is a PhD student in the Centre for Robotics Research at King's College London. He is interested in deep learning, medical imaging, and generative adversarial networks. His current research focuses on developing reliable COVID-19 classification system using deep learning.

H.K. Lam (M'98-SM'12-F'20) received the B.Eng. (Hons.) and Ph.D. degrees from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in 1995 and 2000, respectively. During the period of 2000 and 2005, he worked with the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University as Post-Doctoral Fellow and Research Fellow respectively. He joined as a Lecturer at King's College London in 2005 and is currently a Reader.

His current research interests include intelligent control, computational intelligence and machine learning. He has served as a program committee member, international advisory board member, invited session chair and publication chair for various international conferences and a reviewer for various books, international journals and international conferences. He was an associate editor for IEEE Transactions on Circuits and Systems II: Express Briefs and is an associate editor for IEEE Transactions on Fuzzy Systems, IET Control Theory and Applications, International Journal of Fuzzy Systems, Neurocomputing and Nonlinear Dynamics; and guest editor for a number of international journals. He is on the editorial board of Journal of Intelligent Learning Systems and Applications, Journal of Applied Mathematics, Mathematical Problems in Engineering, Modelling and Simulation in Engineering, Annual Review of Chaos Theory, Bifurcations and Dynamical System, The Open Cybernetics and Systemics Journal, Cogent Engineering and International Journal of Sensors, Wireless Communications and Control. He was named as a highly cited researcher and is an IEEE fellow.

He is a coeditor of two edited volumes: Control of Chaotic Nonlinear Circuits (World Scientific, 2009) and Computational Intelligence and Its Applications (World Scientific, 2012), and author/coauthor of three monographs: Stability Analysis of Fuzzy-Model-Based Control Systems (Springer, 2011), Polynomial Fuzzy Model Based Control Systems (Springer, 2016) and Analysis and Synthesis for Interval Type-2 Fuzzy-Model-Based Systems (Springer, 2016).

Guangyu Jia received the B.S. and M.S. degrees from School of Mathematics, Shandong University, Jinan, China, in 2014 and 2017, respectively. She is currently working toward the Ph.D. degree in The Centre for Robotics Research, Department of Engineering, King's College London, London, UK. Her research interests include machine learning and control theory.