Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of King Saud University – Computer and Information Sciences 34 (2022) 6199–6207

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Attention mechanism and mixup data augmentation for classification of COVID-19 Computed Tomography images

Özgür Özdemir [a,*], Elena Battini Sönmez [a]

[a] Computer Engineering Department, Istanbul Bilgi University, Turkey

ARTICLE INFO

Article history:
Received 22 May 2021
Revised 1 July 2021
Accepted 7 July 2021
Available online 15 July 2021

Keywords:
COVID-19
Classification
Computed Tomography (CT) images
Mixup
Data augmentation
Attention

ABSTRACT

The Coronavirus disease is quickly spreading all over the world and the emergency situation is still out of control. Latest achievements of deep learning algorithms suggest the use of deep Convolutional Neural Network to implement a computer-aided diagnostic system for automatic classification of COVID-19 CT images. In this paper, we propose to employ a feature-wise attention layer in order to enhance the discriminative features obtained by convolutional networks. Moreover, the original performance of the network has been improved using the mixup data augmentation technique. This work compares the proposed attention-based model against the stacked attention networks, and traditional versus mixup data augmentation approaches. We deduced that feature-wise attention extension, while outperforming the stacked attention variants, achieves remarkable improvements over the baseline convolutional neural networks. That is, ResNet50 architecture extended with a feature-wise attention layer obtained 95.57% accuracy score, which, to best of our knowledge, fixes the state-of-the-art in the challenging COVID-CT dataset.

1. Introduction

On January 30th, 2020, the World Health Organization has declared COVID-19 has a Public Health Emergency of International concern (WHO, 2021b). Coronaviruses are a large group of viruses consisting of a nucleus of genetic material surrounded by an envelope and protein spikes, which gives to the virus the shape of a crown or corona. Some types of coronavirus give mild respiratory disease, however other types of virus can generate severe diseases. The 2019 novel coronavirus (2019 NCOV) or COVID-19 was first identified in China and quickly spread all over the world. Common mild symptoms of COVID-19 are temperature, cough, shortness of breath while more severe cases presented pneumonia, kidney failure and death. The website of the World Health Organization (WHO, 2021a) daily updates the numbers: at June 30, 2021, the world's sum of confirmed cases were 181.344.224 and the confirmed deaths 3.934.252.

Currently, campaigns of vaccination are promoted by several countries, but the numbers of vaccines is not enough for the entire world population and the best strategy to control the disease is still to identify and isolate infected people. Coronavirus can be diagnosed by the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test; however, in many realities, the absence or the very limited number of available RT-PCR test kits results in the impossibility to make timely diagnosis to suspected cases, who will carry on spreading the disease, unconsciously (WHO, 2021a). In addition, RT-PCR cannot exactly determine the severity of the disease (Jiang et al., 2020); that is, empirical results showed that the sensitivity of chest Computed Tomography (CT) is higher than the RT-PCR test (Fang, 2020).

As alternative, high resolution CT scans can be successfully used to evaluate the acuteness of COVID-19 and can support doctors to track disease transformation during the follow up (Jiang et al., 2020; Pan et al., 2020; Li and Xia, 2020). That is, recent studies in the medical fields, while confirming the necessity to have early and accurate diagnosis of COVID-19 to reduce the number of fatalities, promote high resolution CT as an essential tool in the diagnosis and follow up of COVID-19 disease. More in details, in the chest CT of patients with severe 2019 NCOV illness it is possible to observe Ground Glass Opacities (GGOs) and patchy consolidation surrounded by GGOs. A successful treatment reduces the size of those spots, and high resolution CTs allow to follow those progresses (Jiang et al., 2020; Pan et al., 2020; Li and Xia, 2020).

* Corresponding author.
E-mail address: ozgur.ozdemir@bilgi.edu.tr (Ö. Özdemir).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.07.005

Important to underline that CTs cannot determine which virus generates pneumonia, however, considering the current situation of emergency, with a high level of spread of COVID-19, CT are practically used in hospitals, and patient with pneumonia are classified as COVID-19 positive (Yang et al., 2020). Moreover, recent advances in deep learning algorithms suggest the use of Computer-Aided Diagnostic (CAD) systems to support front-line doctors for efficient classification of COVID-19 with CT images.

Overall, while several studies have already been made to propose a reliable CAD systems for COVID-19, the contribution of the attention mechanisms have not been analyzed, yet. This work aims to fill in that gap.

Main contributions of this paper are: (1) to propose an ad hoc network for medical images, that exploits an additional feature-wise attention layer connected to the convolution networks; (2) to compare the performance of the proposed feature-wise attention layer against the stacked attention architecture; (3) to achieve state-of-the-art accuracy on the publicly available COVID-CT dataset, using the proposed model and the mixup data augmentation strategy.

The rest of the work is organized as follows. Section 2 overviews the previous works on deep learning for classification of medical images, with special attention to COVID-19. Section 3 details the proposed approach by describing the used deep neural network architecture, introducing the attention mechanism and the mix-up data augmentation technique utilized in this study. Section 4 presents the experimental results and discussions. Finally, conclusion are drawn in Section 5.

2. Previous work on deep learning for COVID-19 classification

Currently, deep networks reach the state-of-the-art performance in many fields, such as image segmentation (Girshick et al., 2014; Long et al., 2015; He et al., 2016; Chen et al., 2018; Biswas et al., 2020; El-Nouby et al., 2021; Yuan et al., 2021; Huang et al., 2021), image classification (Krizhevsky et al., 2012; Sermanet et al., 2013; He et al., 2016; Huang et al., 2017; Zhai et al., 2021; El-Nouby et al., 2021; Yuan et al., 2021; Huang et al., 2021), object detection (He et al., 2016; Lin et al., 2017; Hou et al., 2017; Zhang et al., 2020; El-Nouby et al., 2021; Huang et al., 2021), image captioning (Vinyals et al., 2017; Karpathy and Fei-Fei, 2017; Rennie et al., 2017; Aneja et al., 2018; Hossain et al., 2019) and human action recognition (Liu et al., 2018; Farrajota et al., 2019; Liu et al., 2021). Also in the medical field, deep networks are the latest advances in many areas of research, ranging from classification of pulmonary ground glass opacity nodules (Wang et al., 2021), to classification of benign and malignant tumors on breast images (Kriti and Agarwal, 2020; Pan, 2020; Tripathi et al., 2021; Shia and Chen, 2021), to multiclass classification of skin lesions (Iqbal et al., 2021; Liu et al., 2020), to detection and characterization of Parkinson (Salazar et al., 2020) and Alzheimer's diseases (Shao et al., 2020).

All mentioned successes suggest that the use of deep learning could help to tackle the pandemic of COVID-19 and several works have already been presented using CT images of 2019 NCOV. To name a few, Li (2020) collected a database of CT scans consisting in 4,352 images from 3,322 patients; they used deep CNN to analyze the dataset and they concluded that deep learning models can be utilized for accurate distinction among COVID-19, pneumonia and other lung diseases. Butt et al. (2020) tackled the 2-class problem (COVID, Non-COVID) considering CT scans from patients with 2019 NCOV, influenza viral pneumonia, and no-infection. The sensitivity and specificity of deep learning models proved to be superior compared to the RT-PCR test. Zheng et al. (2020) used a pre-trained U-net to segment the 3D CT chest scans, which were then fed into a 3D CNN to give the probability of COVID-19 infection. Gozes et al. (2020) utilized 2D and 3D deep learning models to classify and monitor the Coronavirus disease; i.e. the proposed Computer-Aided Diagnostic systems detects 2019 NCOV with high accuracy and measures the progress of the illness during the time. Song et al. (2020) developed an accurate computer-aided COVID-19 diagnosis system, which was trained and tested using chest CT scans; more in details, the images were collected from 88 patients affected by the Coronavirus disease, 101 patients diagnosis with bacteria pneumonia, and 86 healthy persons. Classification results were promising; moreover, the model could be used to localize ground-glass opacity (GGO), to support front-line doctors. Wang et al. (2020) modified existing deep learning models to deal with chest CT scans on a newly created database of 1,065 CT images, divided into COVID-19 cases (325 images) and typical viral pneumonia (740 images). Mostafiz et al. (2020) proposed a novel architecture to detect COVID-19 from chest X-ray images using deep CNN and discrete wavelets transform for features extraction.

All these works are important witness of the potentials of deep learning in the medical field; unfortunately, most of the used databases are not publicly available. This issue was addressed by He et al. (2020) who created an open access dataset of COVID-19 Computed Tomography (CT) images, namely COVID-CT. The database includes chest CTs from 216 COVID-19 patients. CT images of this database were retrieved from preprints on COVID-19 published into medRxiv, bioRxiv and other Internet sites distributing unpublished preprints on health. By augmenting COVID-CT dataset with images from the Lung Nodule Analysis (LUNA, 2016) dataset, He et al. (2020) applied transfer learning on DenseNet architecture (Huang et al., 2017) and achieved (F1, AUC, accuracy) scores of (0.85, 0.94, 0.86), respectively.

Important to underline that, since the pictures in COVID-CT dataset come from different preprints, there are several disturbance factors such as size, resolution, illumination, contrast, which increase the complexity of the system. Yang et al. (2020) tackled this problem by using the degraded CT pictures only for training; original CT images donated by the Tongji Hospital, Wuhan, China were used for both validation and test. More in details, they manually extracted the CTs pictures and labeled them by reading the captions of the images; pictures storing multiple CT images were manually separated into individual CT. As results, the dataset stores (349, 397) CT images of (COVID, Non-COVID) pictures. The performance achieved by their contrastive self-supervised learning (Hadsell et al., 2006) and transfer learning approach (F1 = 0.90, AUC = 0.98, and accuracy = 0.89) has been judged as "good enough" by senior radiologists of the hospital. Finally, Ahuja et al. (2020) conducted another study on COVID-CT dataset by proposing to apply data augmentation techniques for mitigating the issue of having a limited number of images. They employed a pretrained ResNet18 network on their augmented dataset, which resulted in 99.8% top accuracy score. However, the comparison between this study and above-mentioned works along with our study would not be fair, because their test set contains fewer samples compared to the currently available set. On the other hand, they utilize their approaches to find the abnormalities caused by COVID from CT images.

3. Methodology

3.1. Deep Convolutional Neural Network

Deep Convolutional Neural Networks (CNN) are biologically inspired 2-dimensional (2D) variations of deep Neural Networks (NN), which, currently, hold the record in performance on several computer vision and image processing tasks (Krizhevsky et al.,

2012; Sermanet et al., 2013; Girshick et al., 2014; Long et al., 2015; Vinyals et al., 2017; Hou et al., 2017; Farrajota et al., 2019; Xu et al., 2020; Zhang et al., 2020). Another advantage of Deep CNN is to be "easy to use" because they automatically decompose the input signal into a set of features, at different level of abstractions, and classify it. In other words, hidden layers are stratum of the network corresponding to features, and deep models have the advantage to make feature extraction and classification at once, overcoming the two-steps procedure of classical machine learning algorithms (LeCun et al., 2015).

The two basic layers of a CNN are the convolutional and the down-sampling layers. While convolutional layers create feature map by convolving 2D filters, i.e. set of weights in 2D, on the top of the image, down-sampling layers extract statistics out of non-overlapping partitions of the input 2D features map, e.g. the max-pool layer returns the max value out of every patch of size $n \times n$. The architecture of a deep CNN depends to the number of hidden layers, i.e. the depth of the model, the number of filters used at every layers, i.e. the number of output channels, the type of special blocks used, e.g. residual blocks with different cardinality, and attention blocks.

Deep CNN started receiving a lot of attention since 2012, when Krizhevsky et al. (2012) won the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) with AlexNet architecture. After that, several architectures have been proposed, such as ConvNet or VGGNet (Simonyan and Zisserman, 2014), the Inception module (Szegedy et al., 2015), Resnet (He et al., 2016), ResNext (Xie et al., 2017), DenseNet (Huang et al., 2017) and AttentionNet (Wang et al., 2017).

Comparing to the previous architectures, the key change in Residual Networks (ResNet) (He et al., 2016) is the use of a residual block, that adds the original input, x, to the output feature, $F(x)$, which is calculated by processing the input x with one or more convolutional layers. That is, while in classical deep CNN each layer is fed with the output of the previous one, when using residual blocks, a layer can feed the next layer and others, which are 2–3 steps ahead in the architecture, i.e. residual blocks are skip connection blocks. Looking at the architecture of the block, it is possible to infer that the aim of the residual block is to learn the difference (or residual) between the output and the input signals. The use of residual blocks solves the two main issues related to deep networks, namely the vanishing gradient and the degradation problem, and results in an increase in performance.

Since 2015, the success of ResNet attracted many researchers who proposed several variations of the original architecture. ResNeXt model (Xie et al., 2017) introduces the concept of "cardinality", which is the NeXt dimension, a new hyperparameter of the deep network. Xie et al. (2017) claims that this new parameter is more effective to add accuracy to the network. ResNeXt (Xie et al., 2017) won the 2nd position in the ILSVRC 2016 competition.

Following the success of residual layer connections, the DenseNet architecture proposed by Huang et al. (2017) advances by carrying residual information and connecting all layers with theirs preceding. Since the information propagates along the depth of the network, the feature extractor channels in all layers can be thinner. The resulting network is more compact compared to residual connection architectures.

This study proposes a new deep network architecture, which uses the above-mentioned convolutional networks as feature extractor backbone and exploits attention mechanisms to intensify the discriminating features of classes. The performances of the extended deep CNN architectures are compared against another attention network, AttentionNet (Wang et al., 2017). The newly introduced attention architecture is described in the next section.

Algorithm 1: Training of the proposed network.

Function *train*($X, y, epoch, \beta$, CNN, FCN, D_f, D_α)**:**
  $h \leftarrow$ Initialize hidden states with random weights in dimension of D_f.
  $\mathtt{Attn}_f, \mathtt{Attn}_h \leftarrow$ Initialize 1-layer perceptron with D_α dimension for features and hidden states attention, respectively.
  Attn$\leftarrow$ Initialize 1-layer perceptron with D_f dimension for attentive features.
  for *e in epoch* **do**
    for X_b, y_b *in minibatch(X, y)* **do**
      if *do_mixup* **then**
        $X_b, y_b \leftarrow apply\_mixup(X_b, y_b, \beta)$
        // As given in Eq. 2
      end
      $f \leftarrow \mathtt{CNN}(X_b)$
      $c \leftarrow attention(f, h, \mathtt{Attn}_f, \mathtt{Attn}_h, \mathtt{Attn})$
      $\hat{y_b} \leftarrow \mathrm{softmax}(\mathtt{FCN}(c))$
      $l_b \leftarrow$ cross-entropy$(y_b, \hat{y_b})$
      Back-propagate the loss l_b through the network
    end
  end
end
Function *attention*($f, h, \mathtt{Attn}_f, \mathtt{Attn}_h$, Attn)**:**
                                  // As given in Eq. 1
  $f \leftarrow \mathtt{Attn}_f(f)$
  $h \leftarrow \mathtt{Attn}_h(h)$
  $h \leftarrow \sigma(f + h)$
  $\alpha \leftarrow \mathrm{softmax}(\mathtt{Attn}(h))$
  return $\alpha * f$
end

3.2. Attention mechanism

Medical images carry discriminating features in narrow regions; in case of Coronavirus disease, CT chest scans are read by expert doctors, who know 'what to look for', i.e. Ground-glass opacities (GGOs), and 'where' to search. The implementation of a computer-aided diagnostic system for automatic classification of COVID-19 CT images must be designed ad hoc to deal with the special issues listed below: when the lesion area is little, working with the entire image is miss-leading, and the situation becomes worse when the wound is near by an edge in the picture, which will create a lot of noise for the classifier, lowering the classification performance. In the deep learning environment, all these issues may be addressed via the attention mechanism.

The recently introduced Residual Attention Networks Wang et al. (2017), namely AttentionNet, is a stacked attention architecture that exploits residual learning. Given that the increasing size would lead to difficulties in learning, Wang et al. (2017) utilized the residual schema to alleviate the problem. Furthermore, AttentionNet incorporates sub attention modules, where each module consists of residual trunk and attention mask branches. The trunk branch, which comprises the residual CNN layer, extracts

features from the set of input, while the attention mask branch emphasizes the significance of the extracted features by applying max-pooling down-sampling and linear interpolation up-sampling operations, consecutively. Residual Attention Networks achieved promising results in photographic images, but the success of the model in medical images has not been studied, yet.

Instead of stacking the attention modules, this work proposes to utilize an extension layer connected to a feature extraction network for attention mechanism as Bahdanau et al. (2015). A CNN component is exploited to create feature mappings. Subsequently, the obtained feature maps are fed into a feature-wise attention layer, along with the hidden states of the mappings, which mimic the recurrence behaviour of Recurrent Neural Networks (RNNs) in order to learn vicinity information. More in details, the context vector, c, is calculated as the attentive features, α_i, with dimension D_α and the feature mappings, f with dimension D_f; moreover, the hidden states, h_i, are utilized by sigmoid function, σ. In formula:

$$h_i = \sigma(h_{i-1} + f) \tag{1a}$$

$$\alpha_i = \frac{\exp(h_i)}{\sum_{k=1}^{D_\alpha} \exp(h_k)} \tag{1b}$$

$$c = \sum_{i=1}^{D_f} \alpha_i f \tag{1c}$$

Consequently, context vectors are fed into a fully connected network with softmax activation in order to create probability distribution of the network predictions. Fig. 1 represents the proposed model. Besides, the pseudocode for tranining the network is given in Algorithm 1.

Comparing to the AttentionNet of Wang et al. (2017), the introduced feature-wise attention network has the big advantage to be architecture-independent since the attention layer is connected to the network implicitly. That is, the proposed model maintains end-to-end training with straightforward implementation by not requiring any interior modifications of the deep CNN.

Furthermore, another advantage of using the proposed attention mechanism by an extension component is in terms of network sizes. That is, despite hidden state and attention weight overheads, the size of the feature-wise attention network is smaller compared to the similar-depth variants of stacked attention networks. Besides, the proposed attention mechanism adds $O(D_f^2.D_\alpha)$ overhead, as given in Eq. 1, to the overall complexity of the network. Similar to the comparison in the network sizes, the time complexity required for feature-wise attention layer is smaller than the stacked attention networks. Table 1 gives details on network sizes FLOP values.

3.3. Data augmentation

Neural Networks need to be trained with a lot of data and the very little dimension of the COVID-CT database makes the data augmentation technique a necessary pre-processing step.

Data augmentation allows to increase the number of training images simply by modifying the existing ones. Common methods for data augmentation are random cropping, horizontal and vertical flipping, rotation, zooming, adding noise, changing lighting condition, etc.

The mixup data augmentation technique has been introduced by Zhang et al. (2017). Mixup builds a new artificial training sample, X, by mixing the pixels of two original pictures, X_1 of class y_1 and X_2 of class y_2. In formula:

$$X = \beta X_1 + (1-\beta)X_2 \tag{2a}$$

$$y = \beta y_1 + (1-\beta)y_2 \tag{2b}$$

where $\beta \in [0, 1]$ is sampled from a beta distribution. Fig. 2 visualizes the mixup operation and samples obtained by this operation.

Inspired by the successful results of Sato et al. working on brain PET images (Sato et al., 2020), we used the mixup data augmentation technique to improve the generalization of the networks. Experimental results highlighted the major improvement in performance of the mixup technique compared to conventional data augmentation approaches.

4. Experiments

4.1. Dataset

This study uses the publicly available COVID-CT dataset collected by He et al. (2020); Fig. 3 displays sample pictures of the dataset and underlines the several disturbance factors present in the database such as low resolution, changes in illumination and low contrast. In order to improve the generalization of the networks, two variations of training set have been prepared using data augmentation techniques. The first experiment followed the traditional approach and images were augmented by vertical flipping and random contrast shift, by a factor randomly selected from the interval $[0.8, 1.0]$. As alternative, in the second experiment, the mixup data augmentation technique has been used on the original training set. In both trials the augmented data resulted in 573 COVID and 702 Non-COVID CT scans. Table 2 details the distribution of (COVID Non-COVID) images in the original and augmented training, validation and test sets. Interesting to underline that, with augmentation, the total number of training images changed from (191 + 234) 425 to (573 + 702) 1275, equivalent to a percentage increase of 300%. As result, while the original database has a total of (425 + 118 + 203) 746 images (57% training, 16% validation, 27% test), the augmented collection has (1275 + 118 + 203) 1596 pictures distributed as (80% training, 7% validation, 13% test) (see Fig. 4).

4.2. Experimental setup

In all experiments, all images were resized to a fixed size, 256×256; pixel values were normalized into the internal [0, 1]; the training and validation sets were split into mini-batches of fixed size of 32. All networks have been trained for 100 epochs and the best weights, which achieved the best accuracy score in the validation set, were stored. The error was calculated using the cross-entropy function and the Adam optimizer was used with a learning rate of 2×10^{-5}. Finally, the mixup value of β (in Eq. 2) was selected randomly from Beta$(0.2, 0.2)$ distribution.

We run several experiments using common convolutional networks and AttentionNet architectures. Network weights were initialized either with random values or with the values of networks pre-trained on ImageNet dataset Krizhevsky et al., 2012. For the AttentionNet structures, the configurations elaborated by Wang et al. (2017) were used; for networks utilizing the feature-wise attention layer, the dimensions of feature mapping and attention weights were set to 512 and 256, respectively. All experiments were conducted on a single Tesla T4 GPU which 100 epoch training takes about 20 min. The network implementations were done in PyTorch.[1] For reproducibility, implementations and weights of trained networks will be shared after the publication of the paper.[2]

[1] https://pytorch.org/.

[2] Our code is available at https://github.com/ozgurozdemir/feature-wise-attention-for-covid-detection.

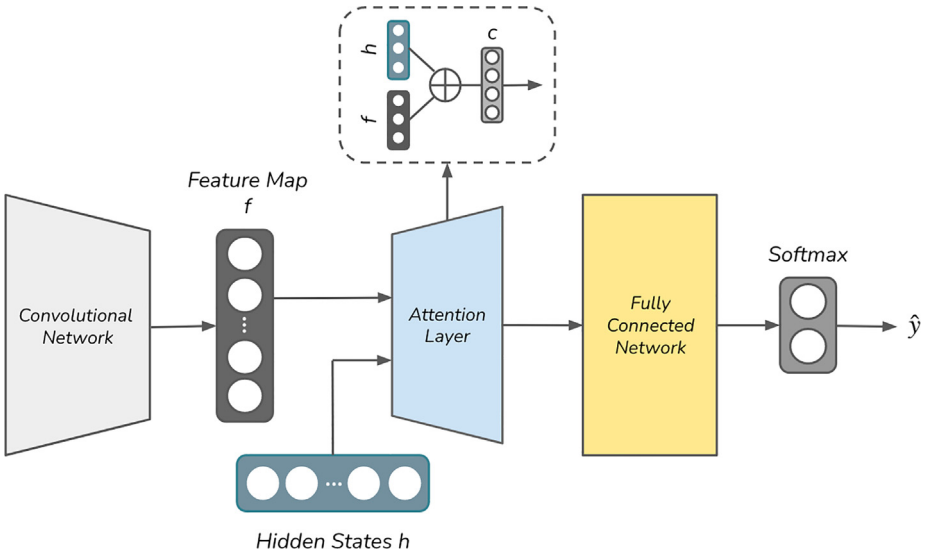


Fig. 1. Proposed network. $f, h, c, \hat{y}$ stands for feature maps, hidden states, context vector and predictions, respectively.

Table 1
Sizes of the network used in this study. Parameters and FLOP quantities are in millions and billions, respectively.

Type	Network	Params(M)	FLOP(B)
Feature-wise Attention	ResNet50	24.69	5.41
	ResNet101	43.68	10.29
	ResNext50 32×4d	24.16	5.61
	Wide ResNet50	68.0	14.97
	DenseNet121	7.6	3.79
	DenseNet201	19.2	5.74
	VGG13	131.1	14.82
	VGG19	141.8	25.71
Stacked Attention	AttentionNet56	29.78	8.32
	AttentionNet92	50.43	13.94

4.3. Results and discussions

Initially, we conducted experiments on different architectures of convolutional neural networks to select possible backbone for our proposed attention mechanism. In these experiments, the network weights were initialized either randomly or with the values of networks pre-trained on ImageNet dataset (Krizhevsky et al., 2012). Table 3 lists the performances of pure convolutional neural networks trained with the original COVID-CT training set, i.e. without augmentation, and tested with the COVID-CT test set. Despite the fact that there was some inconsistencies, the table indicates that initializing weights with pre-trained values improves the performance.

Considering the results of Table 3 the successful networks have been selected to be used as backbone feature extractor in our proposed attention architecture. Table 4 compares the performance of pure convolutional networks and their variations, which utilizes feature-wise attention layer. Looking at the table, it is possible to infer that the addition of the attention layer always increase the initial performance of the models. Moreover, the table proves that the proposed feature-wise attention layer outperforms the stacked attention opponent AttentionNet for each architecture. Eventually, ResNet50 + Attention model achieved the best performance compared to other networks.

Furthermore, in order to improve the generalization of the networks, conventional and mixup data augmentation techniques have been applied to the training set. Table 5 compares the increase in performance of these techniques on both feature-wise and stacked attention networks. Results reported in Table 5 allow to underline that the majority of the models benefit from the mixup data augmentation technique.

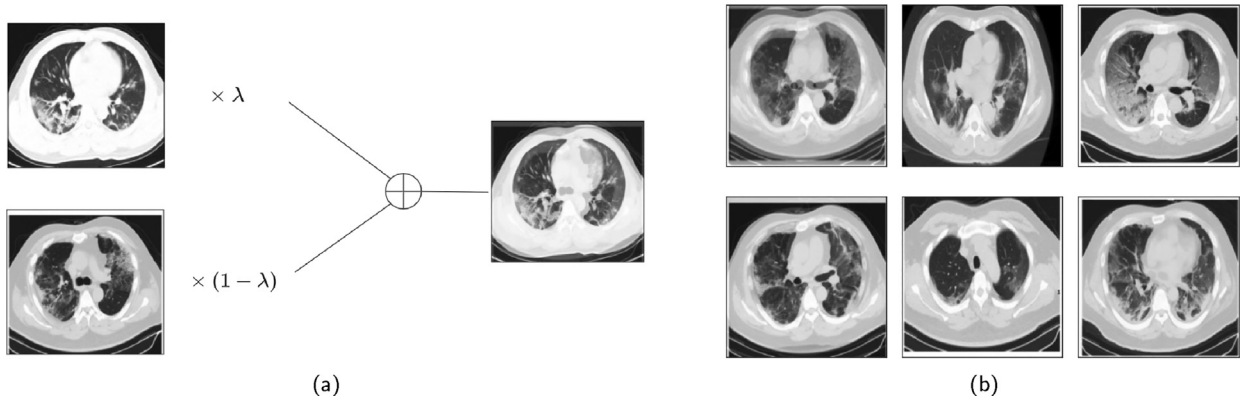


Fig. 2. Mixup data augmentation method. (a) Mixup operation. (b) Samples augmented by mixup from training set.

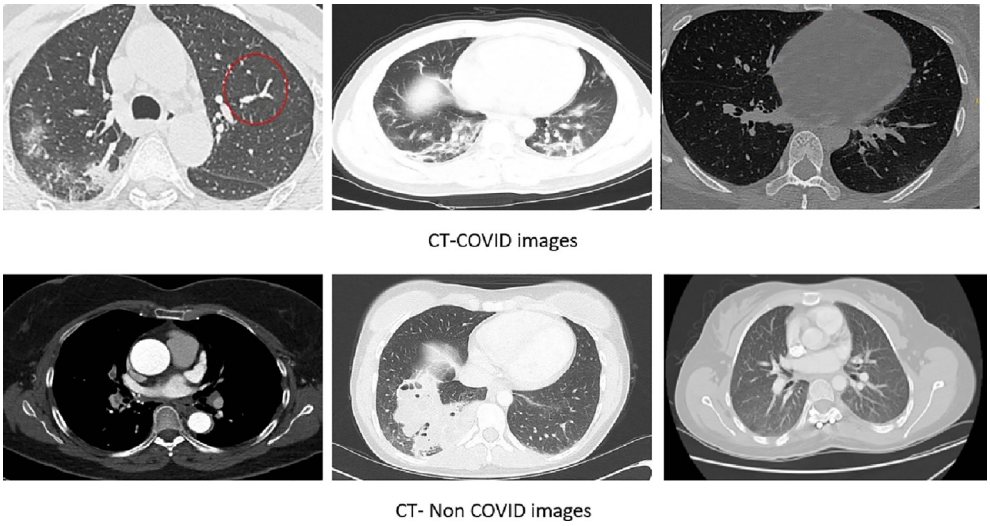


Fig. 3. Images from the COVID-CT database (He et al., 2020). Top row: CT scans with COVID-19; Bottom row: CT scans of h.ealthy patients.

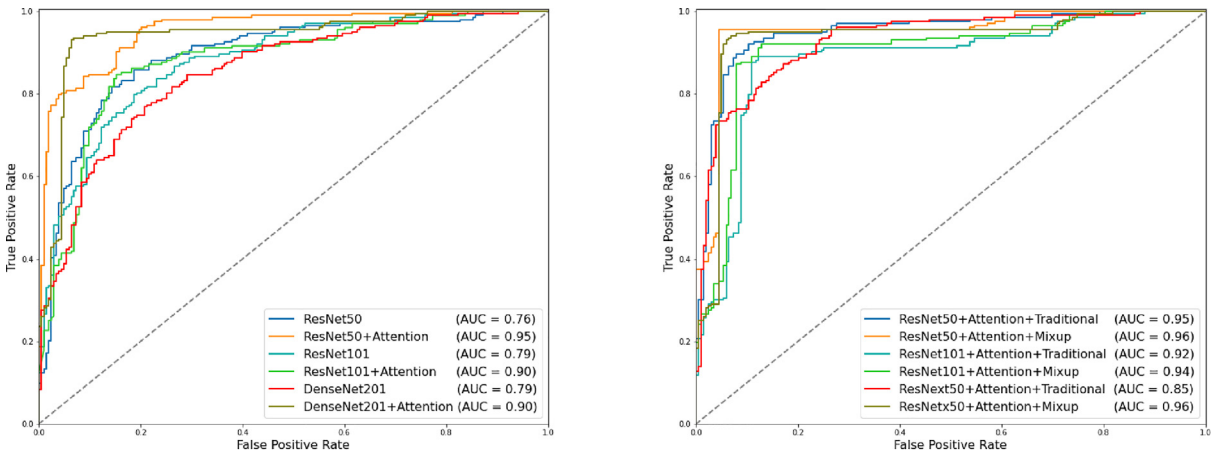
Table 2
Distribution of images in COVID-CT dataset (He et al., 2020).

Set	COVID	Non-COVID
Training	191	234
Training (w/ aug.)	573	702
Validation	60	58
Test	98	105

Besides the experiments assessed on classification performances, an ablation study has been conducted to observe the contribution of the individual components on the proposed attention architecture. Although Table 4 highlights comparison between pure convolutional networks and the proposed approach, attention network contains auxiliary layers, e.g. hidden states (Fig. 1), that may influence the results. In order to develop a fair comparison, a full architecture (Backbone + Attention) has been employed, but the individual component weights have been frozen. As Table 6 indicates, the contribution of the feature extractor stage to the result is major, while feature-wise attention layer is the second significant component. Lastly, the least significant performance drop is obtained by freezing hidden layer, which proves the auxiliary effect of the hidden states component.

Considering previous works on the COVID-CT database, He et al. (2020) used the newly designed transfer learning strategy to reach the top accuracy of 86% on the COVID-CT database, augmented by LUNA's images (LUNA, 2016). Although it is not possible to compare our results against the ones of Yang et al. (2020) because they used the COVID-CT images only for training and the original CT scans employed in the validation and test sets are not publicly available, they achieved 89% top accuracy score. Another study conducted by Ahuja et al. (2020) achieved 99.4% accuracy and 99.6% F1 score, however, because the distribution of training and test set was different compared to the recently available dataset, their networks were tested with fewer samples.

In summary, ResNet50 and ResNext50 32 × 4d architectures with extension of feature-wise attention layer and training set augmented with the mixup technique obtained the best performance of 95.57% accuracy, which is the best result reached on the newly released COVID-CT database.



(a) (b)

Fig. 4. ROC characteristic curves for models trained (a) without data augmentation (b) with data augmentation.

Table 3
Performance of the convolutional networks; all trained without data augmentation. Checkmarks indicates that the networks' weights were initialized with values pre-trained on ImageNet. The weights of of the other networks were initialized randomly. *CE stands for cross-entropy loss.

Network	Pre- training	CE*	Accuracy	F1	AUC	Specificity	Sensitivity
ResNet50		0.63	0.75	0.74	0.75	0.74	0.76
	✓	0.61	0.76	0.73	0.76	0.85	0.67
ResNet101		0.66	0.64	0.60	0.64	0.72	0.55
	✓	0.65	0.79	0.78	0.79	0.78	**0.80**
ResNext50 32×4d		0.66	**0.83**	**0.82**	**0.83**	0.88	0.79
	✓	0.61	0.78	0.79	0.78	0.74	0.83
Wide ResNet50		1.05	0.68	0.66	0.68	0.70	0.65
	✓	0.87	0.75	0.75	0.75	0.70	**0.80**
DenseNet121		0.86	0.75	0.75	0.68	0.74	0.77
	✓	0.85	0.76	0.76	0.76	0.77	0.76
DenseNet201		0.57	0.79	0.76	0.79	**0.89**	0.69
	✓	0.68	0.79	0.77	0.79	0.84	0.73
VGG13		0.74	0.73	0.70	0.73	0.83	0.63
	✓	0.56	0.80	0.80	0.80	0.81	**0.80**
VGG19		0.61	0.71	0.71	0.71	0.70	0.72
	✓	0.60	0.80	0.79	0.80	0.81	0.79
SqueezeNet		0.87	0.71	0.69	0.71	0.75	0.66
	✓	0.63	0.73	0.70	0.73	0.80	0.65

Table 4
Performance of the networks 'with' or without' feature-wise attention layer. All networks were trained without data augmentation.

Network	Attention Layer	CE	Accuracy	F1	AUC	Spec.	Sens.
ResNet50		0.61	0.76	0.73	0.76	0.85	0.67
	✓	0.44	**0.95**	**0.95**	**0.95**	**0.90**	**1.00**
ResNet101		0.65	0.79	0.78	0.79	0.78	0.80
	✓	0.47	0.90	0.91	0.90	0.81	1.00
ResNext50 32×4d		0.66	0.83	0.82	0.83	0.88	0.79
	✓	0.47	0.83	0.82	0.83	0.88	0.79
DenseNet201		0.68	0.79	0.77	0.79	0.84	0.73
	✓	0.57	0.90	0.91	0.90	0.81	1.00
VGG19		0.60	0.80	0.79	0.80	0.81	0.79
	✓	0.65	0.81	0.80	0.81	0.83	0.79
AttentionNet-56		0.61	0.66	0.71	0.66	0.59	0.73
AttentionNet-92		0.63	0.70	0.64	0.69	0.77	0.68

Table 5
Results of the networks utilizing feature-wise attention layer. All networks were trained with data augmentation. * (T) stands for traditional, (M) stands for mixup data augmentations. ** Gain indicates the contribution of using mixup technique instead of traditional augmentations, and it is calculated by mean of the metric scores.

Network	Data Aug.*	CE	Accuracy	F1	AUC	Gain**
ResNet50	(T)	0.5737	0.9507	0.9515	0.9523	
	(M)	0.6281	**0.9557**	**0.9561**	**0.9571**	0.50% ↑
ResNet101	(T)	0.5737	0.9212	0.9245	0.9238	
	(M)	0.6274	0.9360	0.9378	0.9381	1.53% ↑
ResNext50 32×4d	(T)	0.4551	0.8522	0.8387	0.8503	
	(M)	0.6059	**0.9557**	**0.9561**	**0.9571**	12.90% ↑
AttentionNet-56	(T)	0.6247	0.6305	0.5856	0.6276	
	(M)	0.5686	0.7340	0.7353	0.7350	19.56% ↑
AttentionNet-92	(T)	0.6197	0.7340	0.7300	0.7344	
	(M)	0.6076	0.7290	0.7090	0.7276	1.49% ↓

Table 6
Contributions of the components on ResNet50 + Attention network.

Frozen Component	Accuracy	F1	AUC	Drop ↓
ResNet50	0.5862	0.6744	0.5963	35.3%
Feature-wise Attention	0.8325	0.7901	0.8265	14.6%
Hidden States	0.9409	0.9423	0.9429	0.015%
No Freeze	0.9557	0.9561	0.9571	

5. Conclusion and future work

This paper uses deep convolutional neural network models to tackle the urgent and challenging issue of COVID-19 classification of CT images. More in details, this work proposes to exploit an attention layer additionally connected to the CNN for enhancing the discriminative power of the features extracted by CNNs. Experimental results empirically proved our assumptions.

Furthermore, this paper contributes to raise attention on the mixup data augmentation technique, which is employed for improving the generalization of the networks. The best model exploits ResNet50 architecture pipelined with feature-wise attention layer and it is trained with mixup augmented data (in short: ResNet50 + Attention + mixup) to achieve the best performance of 95.57% accuracy. In the baseline study, He et al. (2020) achieved 86% top accuracy score; on the different dataset configurations, Yang et al. (2020) and Ahuja et al. (2020) reached 89% and 99.4% top accuracy performances. With the best of our knowledge, the proposed model, ResNet50 + Attention + mixup, sets the state-of-the-art on the challenging COVID-CT dataset.

Future work includes more experiments related to the feature-wise attention layer to test its robustness against other architectures and types of data. Moreover, we plan also to test alternative techniques for enhancing attention on medical data, such as the method proposed by Srivastava et al. (2015), which uses convolutional networks with a gating mechanism similar to Long-Short-Term-Memory. However, adapting the gating mechanism without any interior modification of the convolutional network, to keep the network architecture-independent, is a challenging task.

Finally, feature-wise attentions could be amplified by using the multi-head attention technique, which achieved significant results in natural language processing domain (Vaswani et al., 2017).

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.

References

Ahuja, S., Panigrahi, B.K., Dey, N., Rajinikanth, V., Gandhi, T.K., 2020. Deep transfer learning-based automated detection of covid-19 from lung ct scan slices. Appl. Intell. https://doi.org/10.1007/s10489-020-01826-w.

Aneja, J., Deshpande, A., Schwing, A.G., 2018. Convolutional image captioning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5561–5570. doi: 10.1109/CVPR.2018.00583.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. http://arxiv.org/abs/1409.0473.

Biswas, B., Ghosh, S.K., Ghosh, A., 2020. A novel CT image segmentation algorithm using PCNN and sobolev gradient methods in GPU frameworks. Pattern Anal. Appl. 23, 837–854. https://doi.org/10.1007/s10044-019-00837-9.

Butt, C., Gill, J., Chun, D., Babu, B., 2020. Deep learning system to screen coronavirus disease 2019 pneumonia. Appl. Intell. https://doi.org/10.1007/s10489-020-01714-3.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., Jegou, H., 2021. Xcit: Cross-covariance image transformers. arXiv:2106.09681.

Fang, Y. et al., 2020. Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology 296, 200–432. https://doi.org/10.1148/radiol.2020200432.

Farrajota, M., Rodrigues, J., du Buf, J., 2019. Human action recognition in videos with articulated pose information by deep networks. Pattern Anal. Appl. 22, 1307–1318. https://doi.org/10.1007/s10044-018-0727-y.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, USA, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P.D., Zhang, H., Ji, W., Bernheim, A., Siegel, E., 2020. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. CoRR abs/2003.05037. https://arxiv.org/abs/2003.05037.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. pp. 1735–1742.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., Xie, P., 2020. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. medrxiv doi: 10.1101/2020.04.13.20063941.

Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H., 2019. A comprehensive survey of deep learning for image captioning. ACM Comput. Surv. 51. https://doi.org/10.1145/3295748.

Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H., 2017. Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3203–3212.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., Fu, B., 2021. Shuffle transformer: Rethinking spatial shuffle for vision transformer. arXiv:2106.03650.

Iqbal, I., Younus, M., Walayat, K., Kakar, M.U., Ma, J., 2021. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. Comput. Med. Imaging Graph. 88,. https://doi.org/10.1016/j.compmedimag.2020.101843 101843.

Jiang, Y., Guo, D., Li, C., Chen, T., Li, R., 2020. High-resolution ct features of the covid-19 infection in nanchong city: Initial and follow-up changes among different clinical types. Radiol. Infect. Diseases 7, 71–77. https://doi.org/10.1016/j.jrid.2020.05.001.

Karpathy, A., Fei-Fei, L., 2017. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. 39, 664–676. https://doi.org/10.1109/TPAMI.2016.2598339.

Kriti, V.J., Agarwal, R., 2020. Deep feature extraction and classification of breast ultrasound images. In: Multimed Tools App. doi: 10.1007/s11042-020-09337-z.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, L. et al., 2020. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology, 65–71. https://doi.org/10.1148/radiol.2020200905.

Li, Y., Xia, L., 2020. Coronavirus disease 2019 (covid-19): role of chest ct in diagnosis and management. Am. J. Roentgenol. 214, 1–7. https://doi.org/10.2214/AJR.20.22954.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944. https://doi.org/10.1109/CVPR.2017.106.

Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G., 2018. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. IEEE Trans. Pattern Anal. Mach. Intell. 40, 3007–3021. https://doi.org/10.1109/TPAMI.2017.2771306.

Liu, L., Mou, L., Zhu, X.X., Mandal, M., 2020. Automatic skin lesion classification based on mid-level feature learning. Comput. Med. Imaging Graph. 84,. https://doi.org/10.1016/j.compmedimag.2020.101765 101765.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2021. Video swin transformer. arXiv:2106.13230.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

LUNA, 2016. Lung nodule analysis 2016, last access: 21.05.2021.https://luna16.grand-challenge.org/data/..

Mostafiz, R., Uddin, M.S., Nur-A-Alam, Mahfuz Reza, M., Rahman, M.M., 2020. Covid-19 detection in chest x-ray through random forest classifier using a hybridization of deep cnn and dwt optimized features. Journal of King Saud University – Computer and Information Sciences. doi: https://doi.org/10.1016/j.jksuci.2020.12.010.

Pan, F., Ye, T., Sun, P., et al., 2020. Time course of lung changes at chest ct during recovery from coronavirus disease 2019 (covid-19). Radiology 259, 715–721. https://doi.org/10.1148/radiol.2020200370.

Pan, X. et al., 2020. Multi-task deep learning for fine-grained classification/grading in breast cancer histopathological images. In: Cognitive Internet of Things: Frameworks. Tools and Applications, Springer, Cham, pp. 85–95.

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1179–1195. https://doi.org/10.1109/CVPR.2017.131.

Salazar, I., Pertuz, S., Contreras, W., et al., 2020. A convolutional oculomotor representation to model parkinsonian fixational patterns from magnified videos. Pattern Anal. Appl. doi:10.1007/s10044-020-00922-4.

Sato, R., Iwamoto, Y., Cho, K., Kang, D.Y., Chen, Y.W., 2020. Accurate bapl score classification of brain pet images based on convolutional neural networks with a joint discriminative loss function. Appl. Sci. 10. https://doi.org/10.3390/app10030965.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. CoRR abs/1312.6229. http://arxiv.org/abs/1312.6229.

Shao, W., Peng, Y., Zu, C., Wang, M., Zhang, D., 2020. Hypergraph based multi-task feature selection for multimodal classification of alzheimer's disease. Comput. Med. Imaging Graph. 80,. https://doi.org/10.1016/j.compmedimag.2019.101663 101663.

Shia, W.C., Chen, D.R., 2021. Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine. Comput. Med. Imaging Graph. 87,. https://doi.org/10.1016/j.compmedimag.2020.101829 101829.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556. http://arxiv.org/abs/1409.1556.

Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., et al., 2020. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. medRxiv https://www.medrxiv.org/content/early/2020/02/25/2020.02.23.20026930.

Srivastava, R.K., Greff, K., Schmidhuber, J., 2015. Training very deep networks. Advances in neural information processing systems, 2377–2385.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Tripathi, S., Singh, S.K., Lee, H.K., 2021. An end-to-end breast tumour classification model using context-based patch modelling – a bilstm approach for image classification. Comput. Med. Imaging Graph. 87,. https://doi.org/10.1016/j.compmedimag.2020.101838 101838.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 5998–6008.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. 39, 652–663. https://doi.org/10.1109/TPAMI.2016.2587640.

Wang, D., Zhang, T., Li, M., Bueno, R., Jayender, J., 2021. 3d deep learning based classification of pulmonary ground glass opacity nodules with automatic segmentation. Comput. Med. Imaging Graph. 88,. https://doi.org/10.1016/j.compmedimag.2020.101814 101814.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., et al., 2020. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). medRxiv https://www.medrxiv.org/content/early/2020/04/24/2020.02.14.20023028.

WHO, 2021a. Coronavirus disease (covid-19), last access: 21.05.2021. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

WHO, 2021b. Who emergencies coronavirus emergency committee second meeting, last access: 21.05.2021. https://www.who.int/.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.

Xu, J., Huang, Y., Cheng, M.M., Liu, L., Zhu, F., Xu, Z., Shao, L., 2020. Noisy-as-clean: Learning self-supervised denoising from corrupted image. IEEE Trans. Image Process. 29, 9316–9329.

Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P., 2020. Covid-ct-dataset: A ct scan dataset about covid-19. CoRR abs/2003.13865. https://arxiv.org/abs/2003.13865.

Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S., 2021. Volo: Vision outlooker for visual recognition. arXiv:2106.13112.

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., 2021. Scaling vision transformers. arXiv:2106.04560.

Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. CoRR abs/1710.09412. http://arxiv.org/abs/1710.09412, arXiv:1710.09412.

Zhang, Z., Gao, J., Mao, J., Liu, Y., Anguelov, D., Li, C., 2020. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. arXiv:2005.04255.

Zheng, C., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., Wang, X., 2020. Deep learning-based detection for covid-19 from chest ct using weak label. medRxiv https://www.medrxiv.org/content/early/2020/03/26/2020.03.12.20027185.