Attention with Multiple Sources Knowledges for COVID-19 from CT Images

Duy M. H. Nguyen,[1, 5] **Duy M. Nguyen,** [2] **Huong Vu,** [3] **Binh T. Nguyen** [4]
Fabrizio Nunnari, [1] **Daniel Sonntag**[1]

[1] German Research Center for Artificial Intelligence, Saarbrücken, Germany
[2] School of Computing, Dublin City University, Ireland
[3] University of California, Berkeley
[4] VNUHCM-University of Science, Ho Chi Minh City, Vietnam
[5] Max Planck Institute for Informatics, Germany

Abstract

Until now, Coronavirus SARS-CoV-2 has caused more than 850,000 deaths and infected more than 27 million individuals in over 120 countries. Besides principal polymerase chain reaction (PCR) tests, automatically identifying positive samples based on computed tomography (CT) scans can present a promising option in the early diagnosis of COVID-19. Recently, there have been increasing efforts to utilize deep networks for COVID-19 diagnosis based on CT scans. While these approaches mostly focus on introducing novel architectures, transfer learning techniques, or construction large scale data, we propose a novel strategy to improve the performance of several baselines by leveraging multiple useful information sources relevant to doctors' judgments. Specifically, infected regions and heat maps extracted from learned networks are integrated with the global image via an attention mechanism during the learning process. This procedure not only makes our system more robust to noise but also guides the network focusing on local lesion areas. Extensive experiments illustrate the superior performance of our approach compared to recent baselines. Furthermore, our learned network guidance presents an explainable feature to doctors as we can understand the connection between input and output in a grey-box model.

Introduction

Coronavirus disease 2019 (COVID-19) is a dangerous infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses 2020). It was first recognized in December 2019 in Wuhan, Hubei, China, and continually spread to a global pandemic. According to statistics at Johns Hopkins University (JHU), until the end of August 2020, COVID-19 caused more than 850,000 deaths and infected more than 27 million individuals in over 120 countries[1]. Among the COVID-19 measures, the reverse-transcription-polymerase chain reaction (RT-PCR) is regularly used in the diagnosis and quantification of RNA virus due to its accuracy. However, this protocol requires functional equipment and strict requirements for testing environments, limiting the rapid and accurate screening of suspected subjects. Further, RT-PCR testing also is reported to suffer from high false-negative rates (Ai et al. 2020). For complementing RT-PCR methods, testings based on visual information as X-rays and computed tomography (CT) images are applied by doctors and have demonstrated effectiveness in current diagnoses, including follow-up assessment and prediction of disease evolution (Rubin et al. 2020). For instance, a hospital in China utilized chest CT for 1014 patients and achieved 0.97 of sensitivity, 0.25 of specificity compared to RT-PCR testing (Ai et al. 2020). Fang et al. 2020 also showed evidences of abnormal CT compatible with an early screening of COVID-19. Ng et al. 2020 conducted a study on patients at Shenzhen and HongKong and found that COVID-19's pulmonary manifestation is characterized by ground-glass opacification with occasional consolidation on CT. Generally, these studies suggest that leveraging medical imaging may be valuable in the early diagnosis of COVID-19.

There have been several deep learning-based systems proposed to detect positive COVID-19 on both X-rays and CT imaging. Compared to X-rays, CT imaging is widely preferred due to its merit and multi-view of the lung. Furthermore, the typical signs of infection could be observed from CT slices, e.g., ground-glass opacity (GGO) or pulmonary consolidation in the late stage, which provide useful and important knowledge in competing against COVID-19. Recent studies focused on three main directions: introducing novel architectures, transfer learning methods, and building up a large scale for COVID-19. For the first category, the novel networks are expected to discriminate precisely between COVID and non-COVID samples by learning robust features and less suffering with high variation in texture, size, and location of small infected regions. For an example, Wang et al. 2020 proposed a modified inception neural network (Szegedy et al. 2015) for classifying COVID-19 patients and normal controls by learning directly on the regions of interest, which are identified by radiologists based on the appearance of pneumonia attributes instead of training on entire CT images. Gozes et al. 2020 and Li et al. 2020 extended the ResNet50 architecture to spot COVID-19 given sequences of chest CT images in 3D dimension. Although these methods could achieve promising performance, the limited samples could potentially simply over-fit when operating in real-world situations. Thus, in the second and third directions, researchers investigated sev-

[1] https://coronavirus.jhu.edu/map.html

eral transfer learning strategies to alleviate data deficiency (He et al. 2020) and growing data sources to provide more large-sized datasets while satisfying privacy concerns and information blockade (Cohen, Morrison, and Dao 2020; He et al. 2020).

Unlike recent works, we aim to answer the question: "*how can we boost the performance of existing COVID-19 diagnosis algorithms by exploiting other source knowledge relevant to a radiologist's decision?*". Specifically, given a baseline network, we expect to improve its accuracy by incorporating two other knowledge: an infected and a heat map region without modifying its architecture. In our settings, infected regions refer to positions of Pulmonary Consolidation Region (PCR) (as shown in Figure 1 at the middle, green regions), a type of lung tissue filling with liquid instead of air; and Ground-Glass Opacity (GGO), an area of increased attenuation in the lung on CT images with preserved bronchial and vascular markings (as depicted in Figure 1 at the middle, red regions). By quantifying those regions, the radiologists can distinguish normal and infected COVID-19 tissues. While infected areas are based on medical knowledge, we refer heat map (as shown in Figure 1 at the right-hand side) as a region extracted from a trained network, which allows us to understand transparently essential parts in the image directly impact the network decision. Our motivation comes from two challenges we observed:

- First, to detect infected COVID-19 patients, clinicians take a look at the high level of CT images. They then examine the local lesion area. Finally, a radiologist can comprehensively consider both global, local information and their prior knowledge as dealing in previous cases to make final judgments. Comparing this idea with learning an autonomous system, while the global features can be derived by training deep networks on entire images, the local and prior knowledge are missing parts in recent works. Thus, we complement it by incorporating infected regions and heat maps and integrating those components during the training process.
- Second, a learning approach using solely global images can suffer from two difficulties. One the one hand, it tends to contain a significant level of noise outside the lesion area. For instance, in figure 1 (right), the lesion area in some conditions can be relatively small (red bounding box) compared with the healthy outside region. Thus, this property makes deep networks hard to focus on the local lesion area and could not be localized precisely positions of disease regions. Furthermore, due to the large inter-class similarity of chest X-ray images, it is challenging to identify the subtle discrepancies of separate classes in the whole images, especially as the critical lesion areas are tiny. By considering these facts, it is crucial to have an attention mechanism to supervise the network such that it can take both lesion regions and global visual information into account for a final decision.

In order to tackle these challenges, this paper introduces a novel fashion to integrate all visual cues via a triplet stream network followed by a fusion branch for diagnosing COVID-19 disease without changing baseline networks' structures.

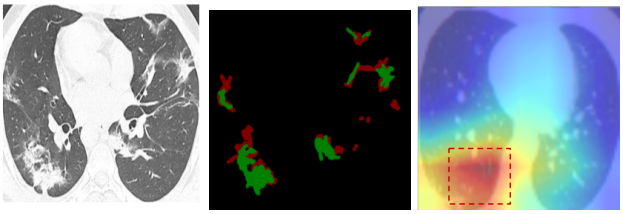

Figure 1: Left: the picture of a COVID-19 case. Middle: red and green labels indicate the Ground-Glass Opacity (GGO) and Pulmonary Consolidation regions (Fan et al. 2020). Right: heatmap region extracted from trained network.

Our architecture is highlighted in two attributes. First, it has two dedicated local branches to focus on local lesion regions, one for infected and another for heatmap areas. In this manner, the influence of the noise in the non-disease areas and missing essential structures can be alleviated. Second, our principal branches, i.e., a global branch and two local branches, are connected by a fusion branch. While the local branches represent the attention mechanism, it may lead to information loss in cases where the lesion areas are scattered in the whole image. Therefore, a global component is demanded to compensate for this error. We reveal that the global and local branches complement each other by the fusion branch, which shows better performance than the current state-of-the-art.

Our contributions can be summarized as follows:

- We provide a new procedure to advance baselines on COVID-19 diagnosis without modifying the network's structures by integrating knowledge relevant to radiologists' judgment as examining a suspected patient. Extensive experiments demonstrate that the proposed method can boost several cutting-edge models' performance, yielding a new state-of-the-art achievement.
- We show the transparency of learned features by embedding the last layer's output vector in the fusion branch to smaller space and visualizing in a 3-D dimension (as shown in Figure 3). Interestingly, the data points in COVID-19 and non-COVID cases can be distinguished with our proposed method. Furthermore, we found a strong connection between input features and network decisions as mapping with activation heatmap and infected regions. Such property is a critical point for clinicians as end-users, as they can interpret how networks create a result given input features in a grey-box rather than a black-box algorithm.

Related Works

In this section, we review two topics that are most relevant to our work, namely *Diagnosis in Chest CT* and *Artificial Intelligence based Applications for COVID-19*.

Diagnosis in Chest CT

In a global effort against COVID-19, the computer vision community pays attention on constructing efficient deep

learning approaches to perform screening of COVID-19 in CT scans. Zheng et al. 2020 pioneered in introducing a novel 3D-deep network (DeCoVNet) composed from pre-trained U-net (Ronneberger, Fischer, and Brox 2015) and two 3D residual blocks. To reduce annotating costs, the authors employed weakly-supervised based computer-aided COVID-19 detection with a large number of CT volumes from the front-line hospital. Other methods also applied 3D deep networks for CT images can be found in (Gozes et al. 2020; Li et al. 2020). In other trends, Song et al. 2020 developed CT diagnosis to support clinicians to identify patients with COVID-19 based on the presence of Pneumonia feature. Shan et al. 2020 and Shi et al. 2020 pursue a strategy of improving accuracy by proposing new architectures such as "VB-Net", infection-size-aware Random Forest (iSARF). To mitigate data deficiency, Xuehai He et al. 2020 builded a publicly-available dataset containing hundreds of CT scans that are positive for COVID-19 and introducing a novelty sample-efficient method based on both pre-trained ImageNet (Deng et al. 2009) and self-supervised learning (Chen et al. 2020). In the same effort, Joseph Paul Cohen et al. 2020 also contributes open image data collection, which was created by assembling medical images from websites and publications.

Artificial Intelligence based Applications for COVID-19

Artificial intelligence has been applied in a large number of treatments against COVID-19 (Dong et al. 2020; Oh, Park, and Ye 2020). Generally, these applications can be divided into three main directions: societal range (e.g., epidemiology Hu et al., 2020), molecular range (e.g., protein structure analytic (Senior et al., 2020)) and patient scale (e.g., medical imaging for diagnosis from CT or X-ray (Wang et al. 2020; Chen et al. 2020)). In this research, we focus on patient range applications (Butt et al. 2020; Shan et al. 2020), especially those based on CT slices. Besides algorithms dedicated to diagnosis COVID-19, deep learning has been employed successfully to segment infection regions in lung CT slices so that the resulting quantitative features can be utilized for different purposes. For instance, Tang et al. (2020) showed that the volume and ratio of infected regions (ground-glass opacity) are positively related to the severity of COVID-19. Furthermore, quantitative features calculated from the right lung are more related to the severity assessment than those of the left lung. Shi et al. 2020 evaluated different ranges of infected lesion sizes to apply for an extensive study of 1658 COVID-19 patients. Rajinikanth et al. (2020) tried to identify the COVID-19 disease by scanning for signs of pneumonia in the lung using CT scans. While recent networks only tackle in a sole target, e.g., only diagnosis or compute infected regions. In contrast, we bring those components into a single system by fusing straight infected areas and global images throughout the learning-network procedure so that these sources can support each other to make our model more robust and efficient. In terms of features, forcing the network to learn on local branches with infected regions can be considered an attention mechanism. Here, the network always takes account of extracted features from these local branches.

Methodology

In this section, we first describe all visual sources used as system input in *Fusion with Multiple Knowledge*, then introducing in detail how to design our architecture and method employed for training in *Network Design and Implementation* subsection.

Fusion with Multiple Knowledge

Infected Branch In Fan et al. 2020, authors developed methods to identify lung areas that are infected by ground-class opacity and consolidation by presenting a novel architecture, namely *Inf-Net*. Its operations are built through a parallel partial decoder used to aggregate the high-level features and generate a global map. Furthermore, by combining the semi-supervised technique, Inf-Net could achieve a state of the art performance on segmentation infected region from CT lung. Given the fact that there is a strong correlation between the diagnosis of COVID-19 and ground-class opacity presented in lung CT scans. We, therefore, adopt the Semi-Infected-Net method from Fan et al. 2020 to localize lung areas suffered by ground-class opacity and consolidation on our CT images. In particular, we expect using this quantification to reduce focused regions of our model to important positions, thus making the system learn efficiently.

Following approach based on semi-supervised data in Fan et al. 2020, we extend it in the diagnosis task by first training the *Inf-Net* on D1 dataset (please see Section Data for further reference). Then, we use this model to obtain pseudo label segmentation masks for 100 randomly chosen CT images from D2 and D3 datasets. After that, we combine the newly predicted masks with D1 as a new training set and re-train our model. The re-trained model will continue to be used for segmenting other 100 ones randomly chosen from the remaining of D2 and D3. Then, we repeated this data combining step. The cycle continues until all images from D2 and D3 have a segmentation mask. We summarize the whole procedure in Algorithm 1.

Algorithm 1: Training Semi-supervised Infected Net

Input: D_{train} = D1 with segmentation masks and D_{test} = D2 $\cup$ D3 without masks.
Output: Trained Infected Net model, M

Set D_{train} = D1; D_{test} = D2 $\cup$ D3; D_{subtest} = NULL
while *len*(D_{test}) > 0 **do**
  Train M
  if *len*($D_{test} > 100$) **then**
    D_{subtest} = random ($D_{\text{test}} \backslash D_{\text{subtest}}$, k = 100)
    $D_{\text{train}} = D_{\text{train}} \cup M(D_{\text{subtest}})$
    $D_{\text{test}} = D_{\text{test}} \backslash D_{\text{subtest}}$
  else
    $D_{\text{subtest}} = D_{\text{test}}$
    $M(D_{\text{subtest}})$
    $D_{\text{test}} = D_{\text{test}} \backslash D_{\text{subtest}}$

Heatmap Branch Besides the whole original scans of CT images, we wanted our proposed network to pay more at-

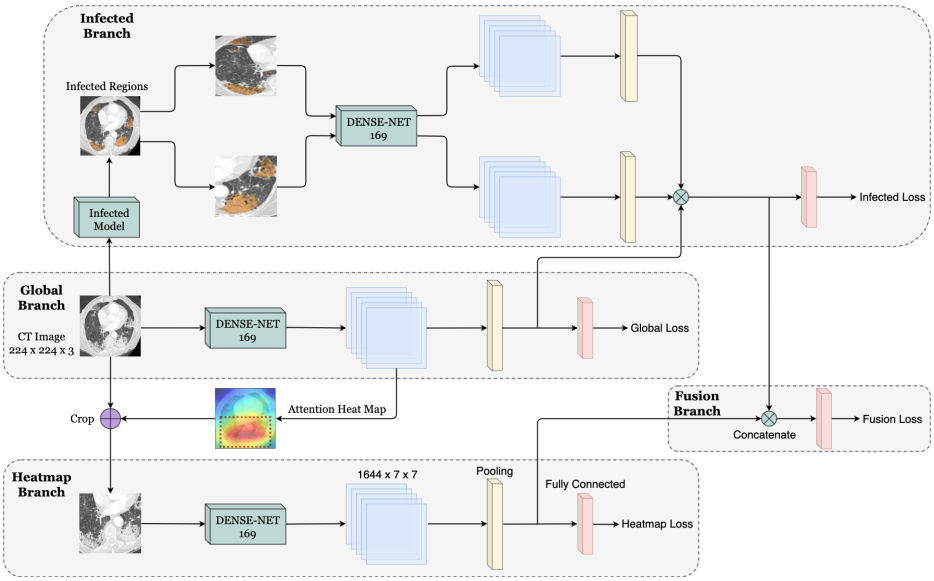

Figure 2: The overview of our triplet stream network with a fusion branch to exploit all features. DenseNet169 is used as an illustration for the baseline network. For all branches, we utilize a binary cross entropy loss function during the training process.

tention to injured regions within each image by building a heatmap branch, which was a separate traditional classification structure as DenseNet169 (Huang et al. 2017) or ResNet50 backbone (He et al. 2016). This additional model was expected to learn the discriminative information from a specific CT scan area instead of the entire image, hence alleviating noise problems.

A lesion region of a CT scan, which could be considered as an attention heat map, was extracted from the last convolution layer's output before computing the global pooling layer of the backbone (DenseNet169 or ResNet50) in the main branch. In particular, with an input CT image, let $f_k(x, y)$ is the activation unit in the channel k at the spatial (x, y) of the last CNN layer, in which $k \in \{1, 2, ..., K\}$ and $K = 1644$ for DenseNet169 or $K = 2048$ for ResNet50 as a backbone. Its attention heat map, H, is created by normalizing across k channels of the activation output by using Eq. 1.

$$H(x, y) = \frac{\sum_k f_k(x, y) - \min(\sum_k f_k)}{\max(\sum_k f_k)} \tag{1}$$

We then binarized H to get the mask B of the suspected region in Eq. 2, where τ is a tuning parameter whose smaller value produces a larger mask, and vice versa.

$$B = \begin{cases} 1, & \text{if } H(x, y) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

We then extracted a maximum connected region in B and mapped with the original CT scan to get the final input for our local branch. A typical example for heatmap region can be seen in Figure 1 at the right-hand side. Given this output and coupling with infected model M obtaining from Algorithm 1, we now have enough input to start training the proposed model.

Network Design and Implementation

Multi-Stream network Our method's architecture can be illustrated in Figure 2, with DenseNet169 as an example of the baseline model. It has three principal branches, i.e., the global and two local branches for attention lesion structures, followed by a fusion branch at the end. Both the global and local branches play roles as classification networks that decide whether the COVID-19 is present. Given a CT image, the parameters of *Global Branch* are first fine-tuned by loading either pre-trained ImageNet or Self-transfer learning tactics as in (He et al. 2020), and continue to train on global images. Then, heatmap regions from the global image extracted using equations (1) and (2) are utilized as an input to train on *Heatmap Branch*. In the next step, input images at the *Global Branch* are fed into Infected-Model M, which is derived after completing the training procedure in algorithm 1, to produce infected regions. Because these lesion regions are relatively small, disconnected, and distributed on the whole image, we find bounding boxes to localize those positions and divide it into two sub-regions: left infected and right infected photos. Those images can be fed into a separate backbone network to output two pooling layers and then concatenating with pooling features from the global branch to train for *Infected Branch*. It is essential to notice that concatenating output features from *Infected Branch* with global features is necessary since, in several cases, e.g., in healthy patients, we could not obtain infected regions. Finally, the *Fusion Branch* can be learned by merging all pooling layers from both global and two local branches.

To be tighter, we assume that each pooling layer is followed by a fully connected layer FC with $C-$ dimensional for all branches and a sigmoid layer is added to normalize the output vector. Let denote $(I_g, W_g, p_g(c|I_g))$, $(I_h, W_h, p_h(c|I_g, I_h))$, and

$(I_{in}, W_{in}, p_{in}(c|I_g, I_{in}))$ as pairs of images, parameters and probability scores belong to the c-th class, $c \in \{1, 2 ..., C\}$ at FC layer for global, heatmap and infected branches, respectively. For fushion branch, we also denote $(Pool_k, W_f, p_f(c|(I_g, I_h, I_{in}))$ as a pair of output feature at pooling layer in branch k ($k \in \{g, h, in\}$), parameter and probability scores belong to the c-th class of the fusion branch.

Then, parameters W_g, W_h, and W_{in} are optimized by minimizing the binary cross entropy loss as follows:

$$L(W_i) = -\frac{1}{C}\sum_{c=1}^{C} l_c \log(\tilde{p}_i(c)) + (1 - l_c)\log(1 - \tilde{p}_i(c)), \tag{3}$$

where l_c is the ground-truth label of the c-th class, C is the total of classes, and $\tilde{p}_i(c)$ is the normalized output network at branch i ($i \in \{g, h, in\}$), which can be computed by:

$$\tilde{p}_i(c) = 1/(1 + \exp(-p_i(c|I_g, I_h, I_{in}) \tag{4}$$

in which

$$p_i(c|I_g, I_h, I_{in}) = \begin{cases} p_g(c|I_g) \text{ if } i = g \\ p_h(c|I_g, I_h) \text{ if } i = h \\ p_{in}(c|I_g, I_{in}) \text{ if } i = in \end{cases} \tag{5}$$

For the fusion branch, we have to compute the pooling fusion $Pool_f$ by merging all pooling values in all branches: $Pool_f = [Pool_g, Pool_h, Pool_{in}]$. After that, we evaluate $p_f(c|(I_g, I_h, I_{in})$ by multiplying $Pool_f$ with weights at FC layer. Finally, W_f can be learned by minimizing equation (3) with formula (4).

Training Strategy Due to the limited amount of COVID-19 CT scans, it is not suitable to train entire all branches simultaneously. We thus followed a strategy that trains each part sequentially to reduce the number of parameters being trained at once. As a branch finished its training stage, its weights would be used to initialize the next branches. Our training protocol can be divided into three stages, as follows:

Stage I: We firstly trained and fine-tuned the global branch, which used architectures from backbones as DenseNet169 or ResNet50. The weight initialization can be done by loading pre-trained ImageNet or Self-Transfer learning method (He et al. 2020).

Stage II: Based on the converged global model, we then created attention heat map images to have the input for the heatmap branch, which was fine-tuned based on the hyper-parameter τ as described in section *Heatmap Branch*. At the same time, we could also train the infected branch independently with the heatmap branch using the pooling features produced by the global model, as illustrated in Figure 2. The weights of the global model were kept intact during this phrase.

Stage III: Once the infected branch and the heatmap branch were fine-tuned, we concatenated their pooling features and trained our final fusion branch with a fully connected layer for the classification. All weights of other branches were still kept frozen while we trained this branch.

The overall training procedure was summarized in Algorithm 2. Different training configurations might affect the performance of our system. Therefore, we analyzed this impact from variation training protocol in the subsection *Performance of Training Strategies*.

Algorithm 2: Training our proposed system

Input: Input image I_g, Label vector L, Threshold τ
Output: Probability score $p_f(c|I_g, I_h, I_{in})$

1 Learning W_g with I, computing $\tilde{p_g}(c|I_g)$, optimizing by Eq. 3 (**Stage I**);
2 Finding attention heat map and its mapped image I_h of I_g by Eq. 2 and Eq. 1.
3 Learning W_h with I_h, computing $\tilde{p_h}(c|I_g, I_h)$, optimizing by Eq. 3 (**Stage II**);
4 Finding infected images I_{in} of I_g by using infected model M;
5 Learning W_{in} with I_{in}, computing $\tilde{p_{in}}(c|I_g, I_{in})$, optimizing by Eq. 3 (**Stage II**);
6 Computing the concatenated $Pool_f$, learning W_f, computing $p_f(c|I_g, I_h, I_{in})$, optimizing by Eq. 3 (**Stage III**).

Experiment and Results

In this section, we present our experimental settings, chosen datasets, and the corresponding performance of different methods.

Data

In our research, we use three sets of data.

- *D1. COVID-19 CT Segmentation from "COVID-19 CT segmentation dataset"*[2].

  This dataset contains 100 axial CT images of more than 40 COVID-19 patients with labeled lung area with ground-class opacity, consolidation and pleural effusion.

- *D2. COVID-19 CT Collection from Fan et al. 2020.*

  This dataset includes 1600 CT slices, extracted from 20 CT volumes of different COVID-19 patients. Since these images are extracted from CT volumes, they do not have segmentation masks.

- *D3. Sample-Efficient COVID-19 CT Scans from He et al. 2020.*

  This data comprises 349 CT images in which 216 of them are from COVID-19 patients. This dataset also does not have segmentation masks and only has COVID-19 positive/negative labels.

For Infected Net model, we exploit all datasets for training. For diagnosis COVID-19, we performed experiments on D3 dataset.

Settings

We implemented several experiments on a TITAN RTX GPU with the Pytorch framework. The optimization used SGD with a learning rate of 0.01 and is divided by ten after 30

[2]https://medicalsegmentation. com/covid19/

Method	Accuracy	F_1	AUC
ResNet50	0.80	0.81	0.88
DenseNet169	0.83	0.81	0.87
Global, Infected, R50	0.83	0.81	0.89
Global, Heatmap, R50	0.82	0.83	0.88
Our Fusion, R50	0.84	0.82	0.91
Global, Infected, D169	0.86	0.83	0.91
Global, Heatmap, D169	0.85	0.82	0.89
Our Fusion, D169	0.87	0.84	0.92

Table 1: Performance between methods using Pre-trained ImageNet. Blue and Red colour are best values for ResNet50 (R50) and DenseNet169 (D169).

Method	Accuracy	F_1	AUC
ResNet50	0.84	0.83	0.91
DenseNet169	0.86	0.85	0.94
Global, Infected, R50	0.84	0.83	0.91
Global, Heatmap, R50	0.87	0.84	0.92
Our-Fusion, R50	0.86	0.87	0.92
Global, Infected, D169	0.85	0.84	0.94
Global, Heatmap, D169	0.87	0.83	0.95
Our Fusion, D169	0.88	0.85	0.96

Table 2: Performance between methods using Self-Transfer. Blue and Red colour are best values for ResNet50 (R50) and DenseNet169 (D169).

epochs. We configured a weight decay of 0.0001 and a momentum of 0.9. For both DenseNet121 and ResNet50 we use batch size of 32 and training for each branch 50 epochs with input size 224×224. The best model is chosen based on early stopping on validation sets. We optimize hyper-parameters τ by grid searching with 0.75, which yields the best performance on the validation set.

Evaluations

In this section, we evaluated our proposed system with different settings and training strategies on the dataset D3. We also compared with state-of-the-art baselines that used ResNet50 and DenseNet169 as backbones (He et al. 2020). The results of both models were taken from the original paper.

Comparing with State of The Art Firstly, from both Table 1 and Table 2, it is clear that our fusion method with ResNet50 and DenseNet169 has significantly improved performance compared to the baseline model in both types of initial weights, from ImageNet and Self-Transfer Learning. More specifically, when using pre-trained ImageNet with ResNet50 backbone, our fusion method increases the accuracy from 80% to 84%, equals the accuracy of the baseline model using ResNet50 with Self-Transfer Learning. Similarly, for DenseNet169, by using pre-trained ImageNet, our fusion method can improve the performance from 83% to 87% in terms of accuracy. This accuracy is even better than the baseline method's accuracy that uses the DenseNet169 backbone and Self-Transfer Learning. The outstanding performance of our fusion method compared to the state-of-the-art is consistent in AUC metric where the fusion method with either DenseNet169 or ResNet50 backbone AUCs topple their baseline AUC, especially using pretrained ImageNet (ResNet50: $91\% > 88\%$ and DenseNet169: $92\% > 87\%$). With Self-Transfer, our fusion method improves the state-of-the-art method's performance, especially with the DenseNet169 backbone. Specifically, fusion's measures increase by 2% in all metrics accuracy, F1, and AUC.

Performance of Mixing Global and Local Branch Using Infected information or Heatmap with the baseline can boost the result by 3%. For example, applying Global-Infected structure for ResNet50 improves the accuracy from 80% to 83%, and the Global-Heatmap network increases the accuracy from 83% to 86% when using pre-trained ImageNet. However, compared to the global models, their improvements from their baselines' results are very similar in all sections. The performance difference is slight, and there is no pattern to conclude if either the Infected or Heatmap branch outperforms the other. Besides, the improvements from both global methods across different backbone methods and types of initial weights are not as remarkable as the fusion method's improvement. This observation strengthens our decision to combine both models in our fusion method to obtain all significant features from each global model.

Training	Global-Infected	Global-Heatmap	Fusion
GHIF	0.82	0.81	0.84
GHI-F	0.83	0.84	0.86
G-H-I-F	0.85	0.87	0.88

Table 3: Accuracy of different training strategies on DenseNet169 with Self-Trans. G: global branch, H: heatmap branch, I: infected branch and F: fusion branch.

Performance of Training Strategies To validate the performance of the proposed multi-step training in Algorithm 2, we evaluated the performance with various strategies: train all branches together (GHIF); train global, heatmap, and infected together and next train fusion branch (GHI-F); and train each branch sequentially (G-H-I-F). The accuracy of these different training methods is presented in Table 3. The table also shows the result of using a distinct combination of branches in our system. The procedure which trained branches together (GHIF and GHI-F) performs worse than training separately (G-H-F-I) with a lower accuracy of about 3%, in which the latter configuration with every single part was trained sequentially performed better other experiments. This phenomenon might be due to the lack of the data as training the whole complex network simultaneously with the limited resources was not a suitable schema. Thus, training each simple branch independently and then fusing them got the best result. Moreover, it was clear that our fusion structure achieved the highest score in all training runs compared to using only two branches. Regarding the joint training setting (GHIF and GHI-F), there were no significant differences in the Global-Infected structure's accuracy with 82% and 83%,

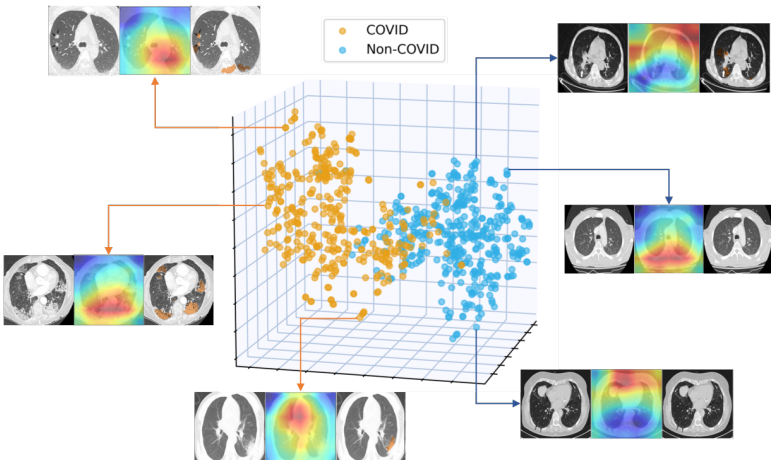

Figure 3: Visualizing learned features by t-SNE with final layers of the fusion branch. Each point is presented together with its original scan, class activation map (CAM) representation, and infected regions (left to right order). For CAM colours, it applies that the closer we get to red in the heatmaps, the stronger the activation is in the original image, which indicates that information from that area contributes strongly to the final decision.

respectively. However, there was an increase of 3% in the scores of the Global-Heatmap network. Both model structures obtained a similar performance with the changes at roughly 1%. The fusion structure also improved 2% between two training strategies. However, it still could surge to 88% in the sequential configuration (G-H-I-F), which was also the highest among all runs.

Visualizing Learned Features

Besides high performance, an ideal algorithm should be explainable to doctors about its connection between learned features and the final decision of network. Such property is critical, especially in medical applications; thereby the reliability is the most concerning factor. To answer this question, we validate our learned features by generating the class activation map (CAM) (Zhou et al. 2016) of the fusion branch and applied t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008) method for visualization by compressing 1644-dimensional features (DenseNet169 case) into a 3D space. Figure 3 depicts the distribution of the pooling features of all images D3 dataset on a 3D plane using t-SNE and CAM representations. Furthermore, infected regions were also shown with their corresponding CT images. The figure presented the COVID-19 and non-COVID points were nearly distinct from others, which confirmed that the features extracted from our fusion part are strongly associated with this disease diagnosis.

By considering CAM color and its corresponding labels, Figure 3 also showed that our system could focus on positions within the lesion lung area for positive scans and vice versa, the red heatmap regions locate outside the lungs for healthy cases. This finding matches the clinical literature that lesion regions inside the lung are one of the major risk factors for the COVID-19 cases (Rajinikanth et al. 2020). Meanwhile, the infected branch also provides useful information by discovering unnormal parts occurring in the lungs (colored in orange). While these lesions are rarely present or appear sparingly in healthy cases, it is clear that this feature plays an important factor in assessing the patient's condition. Finally, given data points that distributed close to the margin separate the COVID-19 and non-COVID cases, the use of other tests to compare results, as well as the experience of the clinician, is a necessary factor in evaluating the actual condition of the patient instead of just relying on the diagnosis of the model. For this property, we once again understand the importance of an explainable model. Without such property, we have a high risk of making mistakes as using automated systems while we could not predict all possible situations.

Conclusion

In this paper, we have presented a novel approach to improve deep learning-based systems for COVID-19 diagnosis. Unlike previous works, we got inspired by considering behaviors of radiologists when examining COVID-19 patients, where all relevant information such as infected regions or heatmaps of injury area are taken into account for the final decision. Extensive experiments show that leveraging all visual cues yields improved performances of two baselines, ResNet50 and DenseNet169, on both pre-trained ImageNet and Self-Transfer initialization. Furthermore, our learned features provides more transparency of the decision process to end-users. As effective treatments are developed, CT images may be combined with additional medically-relevant and transparent information sources. In future research, we will continue to investigate this in a large-scale study to improve performance of proposed system towards explainability as an inherent property of the model.

References

[1] Ai, T.; Yang, Z.; Hou, H.; Zhan, C.; Chen, C.; Lv, W.; Tao, Q.; Sun, Z.; and Xia, L. 2020. Correlation of chest CT and RT-PCR testing in coronavirus disease

2019 (COVID-19) in China: a report of 1014 cases. *Radiology* 200642.

[2] Butt, C.; Gill, J.; Chun, D.; and Babu, B. A. 2020. Deep learning system to screen coronavirus disease 2019 pneumonia. *Applied Intelligence* 1.

[3] Chen, J.; Wu, L.; Zhang, J.; Zhang, L.; Gong, D.; Zhao, Y.; Hu, S.; Wang, Y.; Hu, X.; Zheng, B.; et al. 2020. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. *MedRxiv* .

[4] Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* .

[5] Cohen, J. P.; Morrison, P.; and Dao, L. 2020. COVID-19 image data collection. *arXiv preprint arXiv:2003.11597* .

[6] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* 5(4): 536.

[7] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

[8] Dong, D.; Tang, Z.; Wang, S.; Hui, H.; Gong, L.; Lu, Y.; Xue, Z.; Liao, H.; Chen, F.; Yang, F.; et al. 2020. The role of imaging in the detection and management of COVID-19: a review. *IEEE reviews in biomedical engineering* .

[9] Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images. *IEEE Transactions on Medical Imaging* .

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

[11] Gozes, O.; Frid-Adar, M.; Greenspan, H.; Browning, P. D.; Zhang, H.; Ji, W.; Bernheim, A.; and 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. *arXiv preprint arXiv:2003.05037* .

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

[13] He, X.; Yang, X.; Zhang, S.; Zhao, J.; Zhang, Y.; Xing, E.; and Xie, P. 2020. Sample-Efficient Deep Learning for COVID-19 Diagnosis Based on CT Scans. *medRxiv* .

[14] Hu, Z.; Ge, Q.; Jin, L.; and Xiong, M. 2020. Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112* .

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

[16] Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Bai, J.; Lu, Y.; Fang, Z.; Song, Q.; et al. 2020. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* .

[17] Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

[18] Ng, M.-Y.; Lee, E. Y.; Yang, J.; Yang, F.; Li, X.; Wang, H.; Lui, M. M.-s.; Lo, C. S.-Y.; Leung, B.; Khong, P.-L.; et al. 2020. Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic Imaging* 2(1): e200034.

[19] Oh, Y.; Park, S.; and Ye, J. C. 2020. Deep learning covid-19 features on cxr using limited training data sets. *IEEE Transactions on Medical Imaging* .

[20] Rajinikanth, V.; Dey, N.; Raj, A. N. J.; Hassanien, A. E.; Santosh, K.; and Raja, N. 2020. Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images. *arXiv preprint arXiv:2004.03431* .

[21] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

[22] Rubin, G. D.; Ryerson, C. J.; Haramati, L. B.; Sverzellati, N.; Kanne, J. P.; Raoof, S.; Schluger, N. W.; Volpi, A.; Yim, J.-J.; Martin, I. B.; et al. 2020. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Chest* .

[23] Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A.; et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792): 706–710.

[24] Shan, F.; Gao, Y.; Wang, J.; Shi, W.; Shi, N.; Han, M.; Xue, Z.; and Shi, Y. 2020. Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655* .

[25] Shi, F.; Xia, L.; Shan, F.; Wu, D.; Wei, Y.; Yuan, H.; Jiang, H.; Gao, Y.; Sui, H.; and Shen, D. 2020. Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. *arXiv preprint arXiv:2003.09860* .

[26] Song, Y.; Zheng, S.; Li, L.; Zhang, X.; Zhang, X.; Huang, Z.; Chen, J.; Zhao, H.; Jie, Y.; Wang, R.; et al. 2020. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *medRxiv* .

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

[28] Tang, Z.; Zhao, W.; Xie, X.; Zhong, Z.; Shi, F.; Liu, J.; and Shen, D. 2020. Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. *arXiv preprint arXiv:2003.11988* .

[29] Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; et al. 2020. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *MedRxiv* .

[30] Zheng, C.; Deng, X.; Fu, Q.; Zhou, Q.; Feng, J.; Ma, H.; Liu, W.; and Wang, X. 2020. Deep learning-based detection for COVID-19 from chest CT using weak label. *medRxiv* .

[31] Zhou, B.; Khosla, A.; A., L.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. *CVPR* .