Detection of Covid-19 Based on Mask Region-based Convolutional Neural Networks Using Chest CT-Scan Images

Hamidreza Rokhsati, Alireza Samadi

Abstract

Effective screening of COVID-19 cases has been becoming extremely important to mitigate and stop the quick spread of the disease during the current period of COVID-19 pandemic worldwide. According to the UNDP (United National Development Program) Socio-Economic program, aimed at the COVID-19 crisis, the pandemic is far more than a health crisis: it is affecting societies and economies at their core. There has been greater developments recently in the chest X-ray-based imaging technique as part of the COVID-19 diagnosis especially using Convolution Neural Networks (CNN) for recognizing and classifying images. However, given the limitation of supervised labelled imaging data, the classification and predictive risk modelling of medical diagnosis tend to compromise. In the work, we present COVID-CT-Mask-Net model that predicts COVID-19 from CT scans. The model works in two stages: first, it detects the instances of ground glass opacity and consolidation in CT scans, then predicts the condition from the ranked bounding box detections. To develop the solution for the three-class problem (COVID, common pneumonia and control), we used the COVIDx-CT dataset derived from the dataset of CT scans collected by China National Center for Bioinformation. Using 5000 chest CT-Scan images, the proposed model achieved an accuracy as high as 97.73%, specificity of 98.22% with the precision of 96.65 %. The Mask R-CNN method is found to be accurate and robust in the detection of COVID-19 from chest CT-Scan images.

Keywords: Computer Vision, Mask RCNN, Covid-19 Detection, Chest CT-Scan Images, Mask Region-based Convolutional Neural Networks.

1. Introduction

COVIDNet-CT [1], which consists of a single feature extractor trained on the COVIDx-CT dataset split, COVNet (augmented ResNet50) [2], and ResNet18 [3] are deep learning COVID diagnostic tools for a three-class problem (COVID vs common pneumonia vs control) from CT images. Because of the greater number of potential false predictions, distinguishing between three classes is a more difficult problem than COVID vs Pneumonia or COVID vs Control. To achieve state-of-the-art [1] accuracy, large amounts of data (about 60K images) are required to train the model, which are frequently unavailable, explaining the demand for various augmentations.

Some publications use a semantic segmentation model, such as U-Net in [4, 5], as a pre-processing step for COVID prediction: its output (mask) is used by the classifier to improve prediction. The advantage of using a segmentation model is that it is capable of explicitly learning and predicting COVID-infected areas. COVID-CT and JCS (Joint Classification and Segmentation) [5] are publicly available for binary classification problems (COVID vs control, COVID vs common pneumonia) [6]. COVID-CT predicts the class by combining lung masks predicted by U-Net with deep image features extracted by DenseNet169 and ResNet50, achieving an overall accuracy of 89 percent on the test data of approximately 350 images. JCS employs a

similar approach, but with additional loss functions at deep layers (multiscale training), achieving a Dice score of 0.783 on test data containing approximately 120K images.

Several reviews have directly compared different feature extractors and models to determine the best one for accuracy: [7, 8, 9]. According to these papers, ResNet50 (+Feature Pyramid Network), ResNeXt, and DenseNet121 produce the highest overall accuracy for CT scan data. At least one recent paper [ARK20] discusses the use of Mask R-CNN for predicting COVID from CT scan segmentation.

Due to the prevalence of radiography (X-rays) data, the majority of COVID deep learning models use it, such as the cutting-edge COVID-Net [10], which has an architecture similar to COVIDNet-CT. Only COVIDNet-CT, COVNet, and ResNet18 [3] use CT scans for a 3-class (COVID vs common pneumonia vs control) rather than a binary (COVID vs control) problem, to the best of our knowledge. This problem is more difficult and realistic, both because COVID and common pneumonia symptoms are similar in many ways, but they manifest differently on CT scans [11, 12, 13], but the differences are subtle. These models have the following flaws: COVIDNet-CT requires a large training dataset with various augmentations and class balancing to achieve state-of-the-art accuracy and COVID sensitivity, whereas COVNet was evaluated on a small dataset (about 500 images), ResNet18 [3] is not publicly available, has a low COVID sensitivity (81.2 percent), and was evaluated on a small dataset (90 images).

We hope to address these shortcomings in this paper by extending the semantic segmentation + classification solution to instance segmentation + classification using Mask R-CNN. The state-of-the-art models in instance segmentation and object detection are Mask R-CNN [14] and Faster R-CNN [15]. Mask R-CNN is a Faster R CNN extension that includes a mask prediction branch at the instance level. In contrast, semantic segmentation models such as Fully Convolutional Network (FCN) [16] and U-Net [17] predict objects at the pixel level. Mask R-CNN distinguishes between instances of the same class by predicting their location (bounding box coordinates) using Region Proposal Network (RPN) and Regions of Interest (RoI). As a result, each predicted object has three properties: bounding box, class, and mask. The model's strength stems from the fact that it constructs samples of data from each image (regional features) to make predictions about the instances. This capitalizes on the scarcity of the training data, and we use this strength to obtain accurate predictions while training with a small sample size. We extend Mask R-ability CNN's to detect objects to making predictions about the entire image by augmenting it with a classification module.

The novelty of our approach to COVID-19 prediction can be summarized in the following way:

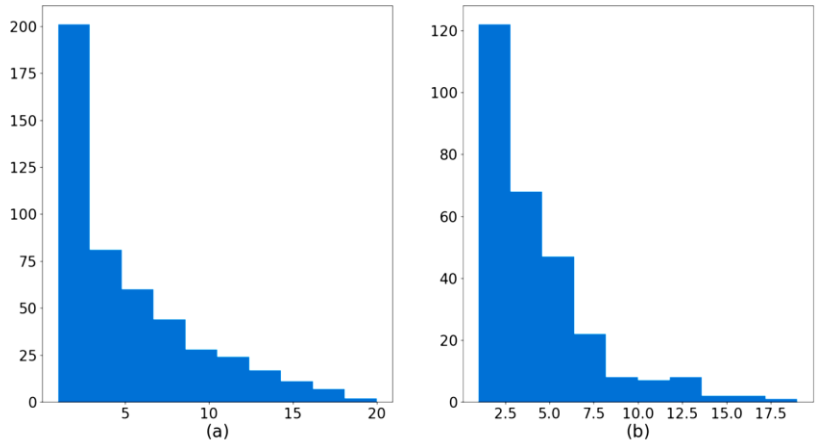
1. Solution: we use approximately 5% of the COVIDx-CT training data and 3% of the total data, and achieve 90.80% COVID sensitivity and 91.66% overall accuracy on the full test split without any data augmentation, such as class weights, background removal, and batch balancing, on which COVIDNet-CT is dependent (21182 images).
2. Architecture: We repurpose Mask R-CNN to predict image class using bounding box predictions by leveraging Mask R-ability CNN's to extract regions of interest (RoIs) from deep features and obtain spatial predictions (bounding boxes) from them to construct a batch of ranked regional predictions in each image and use it to predict global (image) class.
3. By training two models, we solve both segmentation and prediction problems. The Mask R-CNN segmentation model predicts and segments cases of Ground Glass Opacity and Consolidation in CT scans, and COVID-CT-Mask-Net uses this model to predict image class.

Overall, we use much less training data than COVIDNet-CT, achieve higher overall accuracy and COVID sensitivity than COVNet[11] and ResNet18[3], and our solution is more generalizable to other datasets.

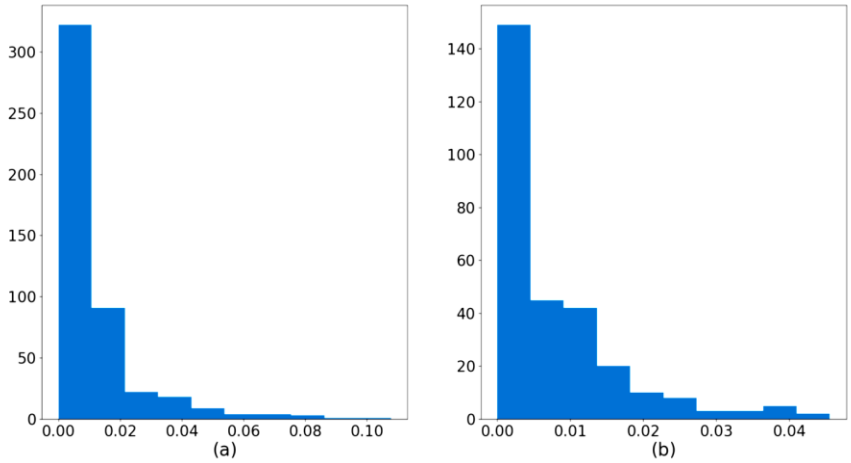
2. Data

2.1. Data Segmentation

We use the publicly available dataset [4] published by China National Center for Bioinformation for our segmentation model, which consists of 5000 scans from 1350 patients with various stages of COVID. At the pixel level, we segmented 1+2 classes: lung field (normal), which we merged with the background, ground glass opacity (GGO), and consolidation (C). Because these two conditions are frequently associated with various stages of COVID and other viral diseases, we classify them as positive. Because we randomly divided the provided dataset into 3000 training, 1000 validation and 1000 testing images while maintaining patient consistency, some slices of COVID-positive patients do not contain positive classes. Figure 1 summarizes the data challenge: it is clear that positive scans can contain a small number of small objects of either class, and overall, the proportion of positive areas is very low, making segmenting them a serious challenge.



(1) Number of instances/image



(2) Ratio of the total area of instances to the image size

Figure 1: COVID correlate distribution in segmentation data. Column (a): Opacity of Ground Glass, Column (b): Consolidation. We plotted the number of distinct GGO and C occurrences in each image in Figure 1.1. The vast majority of images have only a few occurrences of each type (no more than five). Figure 1.2 histograms supplement this finding with the area of occurrences: the absolute majority of them are very small: GGO are 2% of the image size and C are 1%. This means that the vast majority of CT scans have only a few minor occurrences.

2.2. Classification

To compare our model to COVIDNet-CT we used the dataset labelled at the image level provided by the same source, [4], http://ncov-ai.big.ac.cn/download and the split, COVIDx-CT that was used to train COVIDNet-CT model (https://github.com/haydengunraj/COVIDNet-CT), which is publicly available. In total 104900 images from the CNCB dataset were partitioned into 60% training, 20% validation and 20% test data. The difference between COVIDx-CT and the source data is that for COVID and pneumonia classes, only scans with observable infected regions were selected [1]. One of the advantages of our model is a small dataset used for training. We extracted randomly 5000 images from COVIDx-CT training data (1000/class), while maintaining the full size of the validation (21036 images) and test (21182 images) for direct comparison. In the validation split, the shares or Normal/Pneumonia/COVID classes are 43%/35%/22%, in the test split they are 45%/35%/20%.

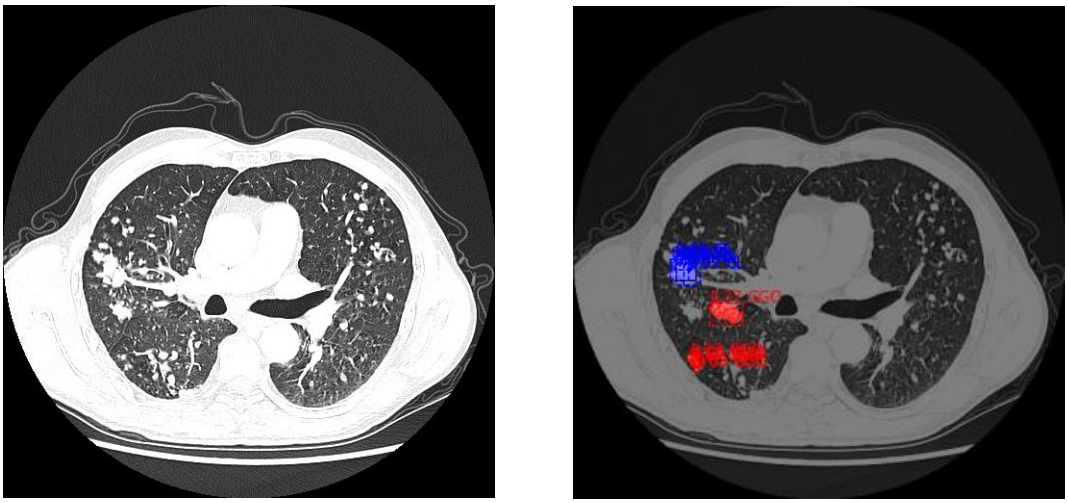


Figure 2: Output of the segmentation model for a lung slice with both Ground Glass opacity and Consolidation for CP

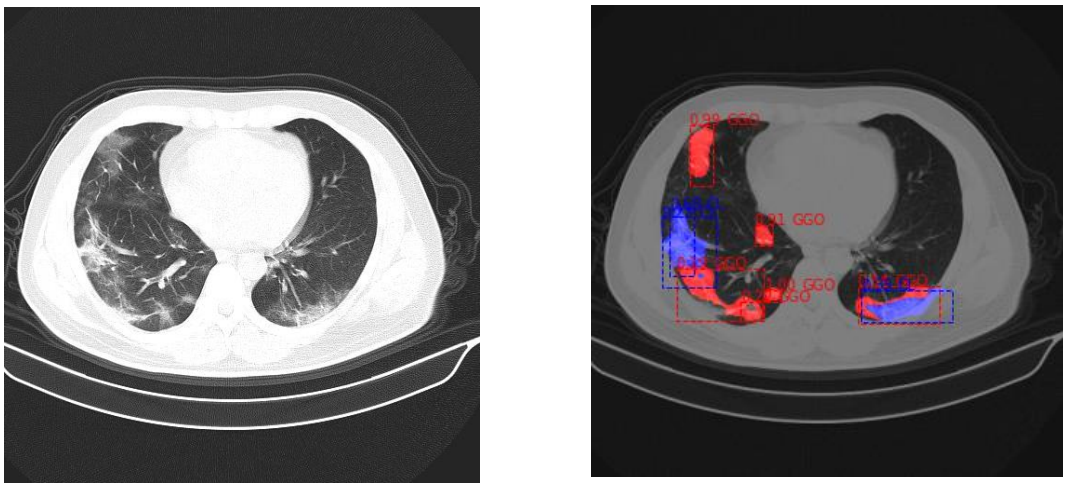
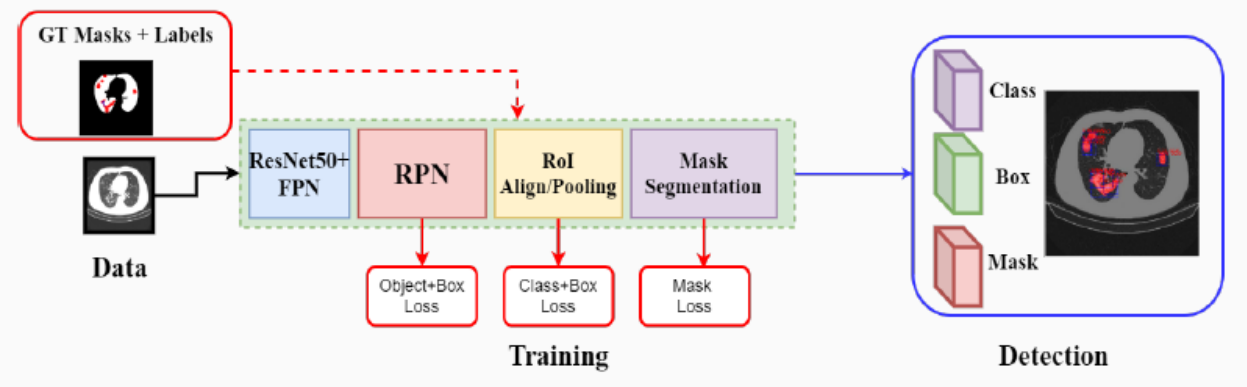


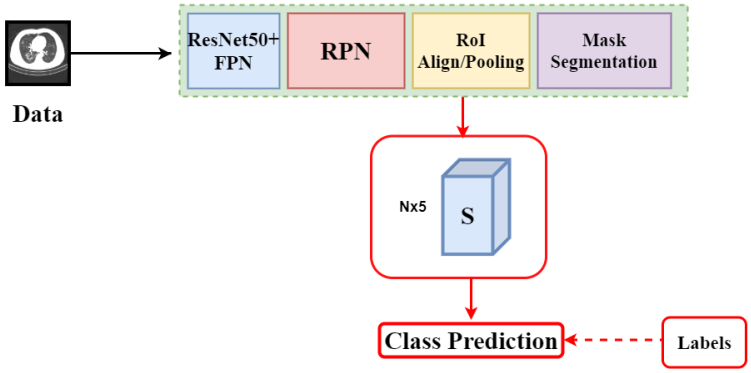
Figure 3: Output of the segmentation model for a lung slice with both Ground Glass opacity and Consolidation for NCP

3. The Proposed Method

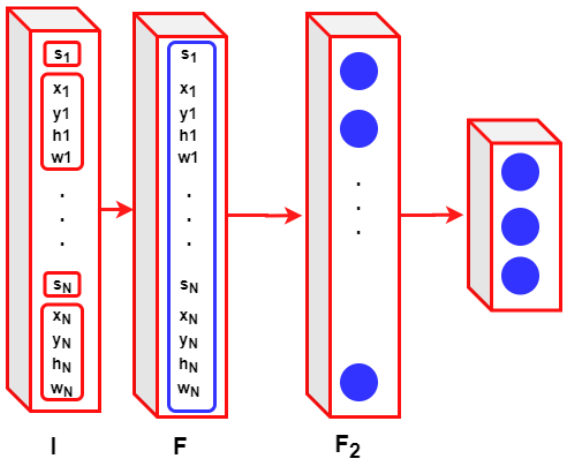
Our solution is divided into two stages: first, we train an instance segmentation model to predict masks of GGO and C areas (Figure 4.1). Following validation, this model is supplemented with a classification module S (Figure 4.3), which classifies the entire input image using ranked bounding box predictions (Figure 4.2).



(1) Mask R-CNN model. Black line is the data input. Labels are the class of the positive objects and bounding box coordinates and binary masks for each instance. Broken red line are the labels and gt masks during training. Red lines are the losses computer during training. Blue line is the output during inference: class, box and mask predictions. The output image is overlaid with the best predictions



(2) COVID-CT-Mask-Net. RPN and RoI do not compute any losses. The new classification module S (Figure 4.3) take the batch size N of the ranked encoded boxes with their scores as an input and predicts the class of the input image.



(3) Classification module S of COVID-CT-Mask-Net: The input I is resized from N * 5 to feature vector F size 1 * N * 5. Fully connected layer F2 is size 1024, and the last prediction layer outpus 3 logit (scores), 1 per image class.

Figure 4: Architecture of the segmentation model, COVID-CT-Mask-Net and classification module S

3.1. Segmentation Model

In three steps, Faster R-CNN [15] and Mask R-CNN [14] extract regional features from one of the backbone feature maps: 1) align the RPN predicted coordinates to the feature map, 2) crop them, and 3) resize (RoI align) to the predefined size. As a result, all RoIs have the same dimensions: C * H *W. (C: number of maps, H, W: height and width of the map). To predict the class and refined coordinates of the object, a sample of positive and negative RoIs is constructed. By comparing the mask's logits to the ground truth mask, the object's mask is predicted independently of other objects and classes. We train Mask R-CNN to build a model capable of detecting a variety of small objects of varying shapes that are common in CT scans of

COVID patients (see Figure 1). The majority of anchor sizes are small (< 32 * 32 pixels) and have a large number of scales (6 total between 0.1 and 2), allowing for accurate detection of various GGO and C shapes. See [14] and our implementation for an explanation of the model's hyperparameters (Non-max suppression, RPN/RoI batch size, foreground and background selection thresholds, and so on). Figure 2 shows some examples of segmentation model outputs. We use Torchvision implementation of Mask R-CNN https://github.com/pytorch/vision/tree/master/torchvision/models/detection with 5 loss functions: binary cross-entropy for class and Smooth1Loss for bounding box coordinates in RPN, multilabel cross entropy for class and Smooth1Loss for bounding box coordinates in RoI and pixel-wise class-conditional binary cross entropy for masks.

Set E:total number of epochs, α: learning rate, λ: weight regularization parameter.
Initialize COVID-CT-Mask-Net with the weights and anchors from the segmentation model.
for 1 **to** E **do**
    Input : Batch of CT images, sparse label vector $\mathbf{L}$ with C classes
    Extract backbone features from the images in the batch
    RPN: predict bounding boxes containing objects and their scores
    RoI: extract N box coordinates predictions and their scores
    Predict N masks (ignored in our implementation)
    Regions Of Interest Output : Batch of N encoded boxes and their confidence scores (tensor $N \times 5$)
    Classifier Module S : accept the ranked boxes and scores, convert batch to feature vector, extract global features
    COVID-CT-Mask-Net Output: Vector of image class predictions $\hat{\mathbf{s}}$
    Binary per-class cross-entropy loss: $\mathcal{L}(\hat{\mathbf{s}}, \mathbf{L}) = -\sum_{k=1}^{C} L_k \times \log \sigma(\hat{s}_k)$
end
Return the best model

Algorithm 1: COVID-CT-Mask-Net algorithm.

R-CNN https://github.com/pytorch/vision/tree/master/torchvision/models/detection with 5 loss functions: binary cross-entropy for class and Smooth1Loss for bounding box coordinates in RPN, multilabel cross entropy for class and Smooth1Loss for bounding box coordinates in RoI and pixel-wise class-conditional binary cross entropy for masks.

3.2. COVID-CT-Mask-Net

We augment Mask R-CNN with a classification module S that makes predictions about the whole image. Details of the COVID-CT-Mask-Net algorithm are presented in Algorithm 1 and Figure 4.2. The details of module S are in Figure 4.3.

Batch to features At the training stage, one of the most important steps in Mask R-CNN is the construction of the batch in the image by taking a sample of positive (score > Θpos) and negative (score Θneg) RoIs. At the inference stage, each RoI predicts a number of encoded bounding boxes (one per class), each with a confidence score, from which a batch of *atmostN* highest-scoring bounding box predictions is extracted after discarding predictions with scores less than score_{Θ} and overlapping predictions. Mask R

CNN encodes coordinates in order to make predictions independent of image size, which is a type of normalization.

We transform this process for the purpose of whole image classification, as we need low-scoring regions too, to give the classifier sufficient information, especially for negative images without any GGO and C conditions. To obtain a fixed-size output from RoI stage, we set the $score_{\Theta}$ = - 0.01, so that even very low-scoring predictions are accepted, and RoI output size is fixed to N * 5 (N encoded bounding box coordinates+confidence score). Normally, RoI decodes these bounding box coordinates by scaling them to the size of the input image, and ranks them based on the confidence score. We ignore the scaling to the image size and use these ranked encoded coordinates with their score as an input in the classification module S. The advantage of this approach is that, even if the highest score is very low (in negative images), the predicted coordinates are still ranked (highest to lowest). This ranking pattern is something the classifier S can learn. S resizes this input into a single feature vector size 1 * (N . 5), which maintains the rank of detections. After some filtering and feature extraction, the module predicts the scores for the whole image (COVID, pneumonia, control), see Figure 4.3.

Non-maximum suppressions (NMS) NMS is the threshold value for discarding predictions of the object of the same class. Setting it high means allowing a larger number of predictions in the training sample with IoU> pre-defined NMS threshold. We established that the model learns that overlapping (adjacent) regions with high scores are associated with higher probability of presence of COVID, and hence it improves sensitivity at the cost of lower overall accuracy. To overcome this fact, since in many scans GGO or C areas can be very small, and hence produce only one or very few high-scoring box predictions, we set the NMS threshold to 0.75 in both models, thus increasing the sensitivity to COVID.

4. Experiments

We re-implement Torchvision's Mask R-CNN library for COVID-CT-Mask-Net. RPN and RoI do not compute any loss during classifier training. The object threshold RoI $score_{\Theta}$ is set to 0.5 in order to accept all box predictions, even those with low scores, and to ensure that the batch size and feature vector remain constant in S. COVID-CT-Mask-Net is trained in three ways: with only classification module S, with module S plus batch normalization layers, and with the entire model. To train the entire model, a large hack was applied to both the RPN and RoI modules: all layers in these modules were set to training mode, the weights were made trainable, and loss computation and all related sampling operations were disabled. Therefore, although formally Mask R-CNN layers were in the evaluation mode, in fact they were updated. Compared to other models, we use a small fraction of the dataset of COVIDx-CT for training, while maintaining the full size of the test and validation sets. As a result, the test/train splits ratio is 7.06, which is the new state-of-the-art, and demonstrates the ability of COVID-CT-Mask-Net to generalize to the unseen data. We use Adam optimizer, learning rate 1e - 5, weight regularization parameter 1e - 3, and train each algorithm for 50 epochs. For other details of the segmentation algorithm and COVID-CT-Mask-Net see our implementation, https://github.com/AlexTS1980/COVID-CT-Mask-Net.

To evaluate each model, we compute the sensitivity/recall and precision/positive predictive value (PPV) for each class C and the overall accuracy of the model:

$$\text{Sensitivity(C)} = \frac{\boldsymbol{True\ Positive(C)}}{\boldsymbol{True\ Positive(C) + False\ Negative(C)}}$$

$$\text{Precision(C)} = \frac{\boldsymbol{True\ Positive(C)}}{\boldsymbol{True\ Positive(C) + False\ Negative(C)}}$$

$$\text{Overall Accuracy} = \frac{\sum_c \boldsymbol{True\ Positive(C)}}{\sum_c \boldsymbol{True\ Positive(C) + \sum_c False\ Negative(C)}}$$

Best results for each trained version of COVID-CT-Mask-Net are presented in Table 1. The model with the classifier head + batch normalization layers produces precision > 90% across all classes. Comparison of our results to other COVID CT detectors for 3 classes is presented in Tables 1 and 2. For COVIDNet-CT we used the best reported model (COVIDNet-CT-A), COVNet and [3] report only one model.

Table 1: Sensitivity (precision) and overall accuracy results on COVIDx-CT test data (5000 images)

CovidNet (ResNet50)	COVID	Pneumonia	Normal	Overall
Accuracy	98.40%	96.60%	98.20%	97.73%
Precision	97.64%	94.64%	97.69%	96.65%
Recall	94.95%	95.72%	98.14%	96.27%
Sensitivity	94.95%	95.72%	98.14%	96.27%
Specificity	99.36%	97.07%	98.24%	98.22%

Table 2: Comparison to other models. The results for COVIDNet-CT were obtained by running the publicly available model (https://github.com/haydengunraj/COVIDNet-CT) on the same test split, results for the other two models are taken from the publication. Last column is the share of COVID observations in the test split. Test split for COVNet has 438 images, ResNet18 90 images.

Model	COVID Sensitivity	Overall accuracy
Ours	96.27%	97.73%
COVIDNet-CT [1]	92.49%	97.57%
COVNet [2]	90.00%	89.04%
ResNet18 [3]	81.30%	86.70%

5. Conclusion

Finding a sufficiently large dataset to train models for accurate COVID predictions is frequently difficult. One of the most powerful aspects of COVID-CT-Mask-methodology Net's is its ability to train on very small amounts of data with no balancing or augmentation adjustments. We trained our model on less than 5% of the COVIDx-CT training split and tested it on over 1000 test images, achieving 97.73% overall accuracy and 96.27% COVID sensitivity. The model can be easily and quickly finetuned to new CT data to achieve high COVID detection rate. The source code with all models and weights are on https://github.com/AlexTS1980/COVID-CT-Mask-Net.

References

1. Hayden Gunraj, Linda Wang, and Alexander Wong. Covidnet-ct: A tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images. arXiv preprint arXiv:2009.05383, 2020.
2. Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology, 2020.
3. Charmaine Butt, Jagpal Gill, David Chun, and Benson A Babu. Deep learning system to screen coronavirus disease 2019 pneumonia. Applied Intelligence, pages 1–7, 2020.
4. Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. Cell, 2020.
5. Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Chao-Wei Zhao, and Ming-Ming Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. arXiv preprint arXiv:2004.07054, 2020.
6. Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865, 2020.
7. Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Huiying Zhao, Yusheng Jie, Ruixuan Wang, Yutian Chong, Jun Shen, Yunfei Zha, and Yuedong Yang. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. medRxiv.
8. Feng Shi, Jun Wang, Jun Shi, Ziyan Wu, Qian Wang, Zhenyu Tang, Kelei He, Yinghuan Shi, and Dinggang Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE reviews in biomedical engineering, 2020.
9. Boyi Liu, Bingjie Yan, Yize Zhou, Yifan Yang, and Yixian Zhang. Experiments of federated learning for covid-19 chest x-ray images. arXiv preprint arXiv:2007.05592, 2020.
10. Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. arXiv preprint arXiv:2003.09871, 2020.
11. Wei Zhao, Zheng Zhong, Xingzhi Xie, Qizhi Yu, and Jun Liu. Ct scans of patients with 2019 novel coronavirus (covid-19) pneumonia. Theranostics, 10(10):4606, 2020.
12. Wei Zhao, Zheng Zhong, Xingzhi Xie, Qizhi Yu, and Jun Liu. Relation between chest ct findings and clinical conditions of coronavirus disease (covid-19) pneumonia: a multicenter study. American Journal of Roentgenology, 214(5):1072–1077, 2020.
13. Tao Yan, Pak Kin Wong, Hao Ren, Huaqiao Wang, Jiangtao Wang, and Yang Li. Automatic distinction between covid-19 and common pneumonia using multi-scale convolutional neural network on chest ct scans. Chaos, Solitons & Fractals, 140:110153, 2020.
14. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.
15. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.
16. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

17. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.