Segmentation of Lungs, Lesions, and Lesion Types on Chest CT Scans of Patients with Covid-19[⋆]

Daria Lashchenova[1][0000−0002−1894−9877], Alexander Gromov[1,3][0000−0001−9818−3770], Anton Konushin[1,2][0000−0002−6152−0021], and Anna Mesheryakova[3][0000−0002−2409−0018]

[1] Lomonosov Moscow State University, Moscow, Russia
{daria.laschenova, alexander.gromov, anton.konushin}@graphics.cs.msu.ru
[2] NRU Higher School of Economics, Moscow, Russia
[3] Third Opinion Platform LLC, Moscow, Russia
{alexander.gromov, ceo}@3opinion.ai

Abstract. The covid-19 pandemic has quickly spread all over the world, overwhelming public healthcare systems in many countries. In this situation demand for automatic assistance systems, to facilitate and accelerate a doctor's job has rapidly increased. Antibody tests were introduced for diagnosing covid-19, but physicians still need tools for quantification of disease severity, since treatment choice strongly depends on it. To estimate the severity of the disease physicians use computer tomography scans. It provides physicians with information about lung lesions and their types and they use this information to determine proper treatment. In this paper we made an attempt to build a system that uses patients' computer tomography scans for lung and lesion segmentation and for segmentation of specific types of lesions (i.e. pulmonary consolidation and "crazy-paving"). Models for lung, lesions, consolidation, and "crazy-paving" segmentation performed with 0.96, 0.65, 0.48, 0.45 Dice coefficients respectively. Also it was shown that removing images with inaccurate ground-truth from the training subset can improve the quality of models trained on it.

Keywords: Covid-19, CT, Lesion segmentation, Lung segmentation, Deep learning

1 Introduction

Covid-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that has a considerable mortality rate. Quick spread of this disease caused a pandemic, which overwhelmed healthcare systems in a large number

of countries. As of 1 July 2020, 10.5M cases were confirmed worldwide and 512 thousands of patients died, so the average mortality rate was about 5%. In Russia it was 1.5% (9.5 thousands deaths for 654 thousands cases).

Early diagnosis can reduce the time and intensity of medical treatment. This can be achieved by the use of computer tomography (CT). CT is a radiography in which a three-dimensional image of a body structure is constructed by computer from a series of plane cross-sectional images made along an axis. CT provides physicians with information about the lesions in lungs. Then they can calculate the percentage of lung opacity and identify the type of the lesion so they can select appropriate treatment for the patient and monitor the course of the disease.

Computed tomography (CT) is considered to be a primary tool for giving diagnosis of covid-19 and evaluation of the disease progression. CT is a radiography in which a three-dimensional image of a body structure is constructed by computer from a series of plane cross-sectional images made along an axis. CT provides radiologists with information about disease features, or radiographic findings. These features are examined in order to identify their type and volume. Then this information is used for selection of appropriate treatment for the patient and monitoring the course of the disease.

The primary CT findings of covid-19 have been reported in [10]. They include pulmonary consolidation, "ground-glass" opacity and "crazy-paving" pattern. Pulmonary consolidation is a region of normally compressible lung tissue filled with liquid instead of air. "Ground-glass" opacity is a descriptive term referring to an area of increased attenuation in the lung on CT scans with preserved bronchial and vascular markings. "Crazy-paving" pattern refers to the appearance of "ground-glass" opacity with superimposed interlobular septal thickening and intralobular septal thickening. The lack of sufficient method for lesion volume estimation and an enormous amount of CT scans for analysis increased demand for supervising systems.

The aim of this study was to create a solution for segmentation of lungs and abnormal regions of lungs and for detecting regions with pulmonary consolidation and "crazy-paving" pattern.

2   Related work

As covid-19 has quickly been spreading, computer vision scientists started to search for solutions that could help physicians diagnose their patients. There were several areas to research.

Before the mass use of antibody tests some scientists would try to find out if a patient had covid-19, using only their CT scans. Linda Wang [9] introduced COVID-Net, a neural network architecture that could classify if a person was ill and distinguish covid-19 from non-covid-19 pneumonia. Xuehai He [2] used CRNet [5] to detect covid-19. Xiaolong Qi [6] estimated how much time the patient would spend in a hospital using only their CT.

But those studies could not be applied for tracking patients' condition. To achieve that, some scientists concentrated on a segmentation task, detecting lungs and areas with lesions. This information could then be used for measuring the percentage of lung opacity to quantify severity of the disease. Segmentation could be performed on 2D

horizontal slices of CT or on full 3D scans. Lu Huang [3] used U-net [7] architecture on horizontal slides of CT scans. Shuo Jin [4] used 3D Unet++ [11] on CTs with a different slice thickness. Fei Shan [8] proposed 3D-model VB-Net for segmentation of lesions, lung lobes and lung segments, training the model with human-in-the-loop strategy.

3 Dataset

The dataset was provided by Third Opinion Platform [1]. It contained 529 studies of patients with and without covid-19. Studies came in dicom format, clipping to the window (according to window height and window width dicom parameters) was not performed.
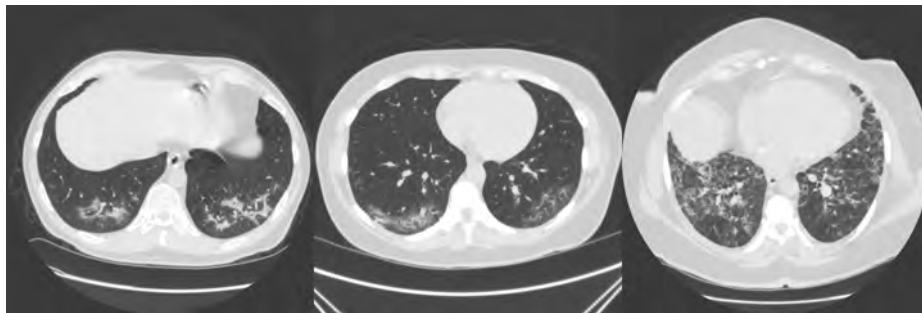

Fig. 1: Examples of horizontal slices of CT.

10-20 horizontal slices from each study were assessed giving us 10454 images. Radiologists assessed areas with lungs, regions of pulmonary consolidation and regions with "ground-glass" opacity, particularly with "crazy-paving".
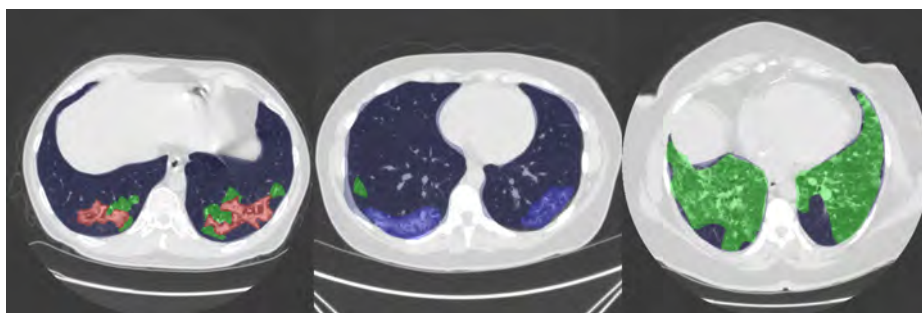

Fig. 2: Examples of horizontal slices of CT.

The dataset was split into training and testing subsets, containing 85% and 15% studies respectively.

Table 1: Dataset statistics.

Class	Number of images with the class	Number of studies with the class
lung	7873	528
lesion	5303	464
"ground-glass"	4990	461
consolidation	489	287
"crazy-paving"	1691	114

Radiologists used an assessment tool to draw polygons around regions of interest. As they tend to draw areas with smoothed boundaries, while models for segmentation calculate more precise masks, it is not expected to get results close to ideal according to quality metrics.

4 Metrics

For model evaluation three metrics were used: mean AP over test studies, IoU and Dice coefficient.

Model's recall is calculated as the number of true positive pixels divided by the number of all positive pixels. Model's precision is calculated as the number of true positive pixels divided by the number of pixels, predicted as positive. AP is calculated as an area under precision-recall curve, where each point on the curve plot represents precision and recall of results with a threshold corresponding to the point.

IoU measures how much ground truth and predicted areas overlap:

$$IoU = \frac{TP}{FN + TP + FP} \tag{1}$$

Dice coefficient is often used to evaluate the quality of segmentation of medical images:

$$Dice = \frac{2TP}{FN + 2TP + FP} \tag{2}$$

TP is the number of true positive pixels, FN – number of false negative pixels, FP – number of false positive pixels.

5 Proposed method

Due to the fact that only 20% horizontal slices from each CT were assessed, it was necessary to use a neural network for 2D segmentation of images. In this work U-net [7] was used. It takes horizontal slices of lung CT and returns probabilities of belonging to a certain class for each pixel.

At first, image normalization is performed on each slice image:

$$I_{norm} = \frac{I_{orig} - \mu}{\sigma} \tag{3}$$

where mu is the mean of the image and $sigma$ is its standard deviation. It was done to unify information from different X-ray machines, as due to different settings they can produce information in different ranges of values. Then random spatial transformations (horizontal flipping, shifting, scaling, rotating) were applied on every slice to augment images.

Segmentation of lungs and regions with lesions. The first task was to segment lesions, so it could be possible to calculate the percentage of lung opacity. Diseased areas and lungs are segmented on each horizontal slice.

Several losses were used to train U-net. For training with binary cross entropy loss and Dice loss models predicted two masks: probability of lung in the pixel and probability of affected lung (either pulmonary consolidation or "ground-glass" opacity). Another model was trained using softmax cross entropy loss. It predicted 3 classes: background, healthy lung and affected lung, so the mask with background was added and lung class was replaced by healthy lung.

In the table 2 results for lung segmentation and lesion segmentation are shown.

Table 2: Results for segmentation of lungs and lesions.

Experiment	mAP (lung)	IoU (lung)	Dice (lung)	mAP (lesion)	IoU (lesion)	Dice (lesion)
BCE loss	**0.99591**	**0.94131**	**0.9694**	0.784	0.51041	0.63645
Dice loss	0.99494	0.94038	0.96898	**0.78404**	**0.52256**	**0.65046**
Softmax CE loss	0.992	0.93689	0.96713	0.78494	0.50343	0.62957

The best result for lesion class by all metrics was shown by the model that was trained with Dice loss. Models with BCE loss and Dice loss showed similar results for lung class.

Segmentation of regions with pulmonary consolidation and "crazy-paving". The second task was to identify different types of pathologies. Experiments showed that the use of samplers was necessary in this task, as the majority of images did not contain consolidation or "crazy-paving", so training was unstable as the model received an enormous amount of negative examples.

Several models were trained. The first one (U4) predicted four binary masks for the following classes: lungs, consolidation, "ground-glass" opacity, and "crazy-paving". During training batch size was divisible by 4. In every four elements of batch the first element contained image with consolidation, the second — "crazy-paving", the third — "ground-glass" opacity and the fourth contained image with healthy lungs with 95% probability and image without lungs (i.e. slices from the top or the bottom of CT) with 5% probability.

The second model (C) predicted two binary masks: mask for lung and for consolidation probabilities. Sampler was used so each sample in a batch during training contained

consolidation with 95% probability, lungs without consolidation with 2.5% probability and no lungs with 2.5% probability. The third model (CP) was similar to the second, but made predictions for "crazy-paving".

All three models were trained with binary cross entropy loss, Dice loss, and softmax cross entropy loss. For training with CE loss, the model was supposed to predicted the background class and unaffected lung class instead of the lung class.

Table 3: Results for consolidation segmentation.

Experiment	mAP	IoU	Dice
U4 BCE	0.51862	**0.42743**	**0.48593**
U4 Softmax CE	0.5064	0.40548	0.47363
C BCE	**0.54077**	0.2652	0.33648
C Dice	0.49091	0.25771	0.32907
C Softmax	0.52354	0.24553	0.31573

Table 4: Results for "crazy-paving" segmentation.

Experiment	mAP	IoU	Dice
U4 BCE	0.3785	**0.42743**	**0.45657**
U4 Softmax CE	0.36422	0.42337	0.44795
CP BCE	**0.386**	0.2231	0.25564
CP Dice	0.3333	0.17061	0.20176
CP Softmax	0.3724	0.1869	0.21889

Tables 3, 4 show results for consolidation and "crazy-paving".

The best result was shown by the united model with four classes for prediction with binary cross entropy loss.

6　Analysis of results

For U4 BCE model confusion matrices for train and test subsets were calculated.

As figure 4 shows, on both train and test subsets the model confuses lesion classes and healthy lungs. This can be explained by the fact that some assessors tend to draw inaccurate masks, while models produce precise result.

Also figure 4 shows, that on both train and test subsets the model significantly confuses "ground-glass" opacities that are not "crazy-paving" and the "crazy-paving". It also confuses consolidation and "ground-glass". This could be explained by the presence of ambiguous examples in the dataset and/or noisy assessment, examples are presented on figure 5. To check the latter hypothesis three models (U4 BCE, C BCE, C Dice) were applied to the train dataset. Images from the train dataset were blocklisted if IoU of
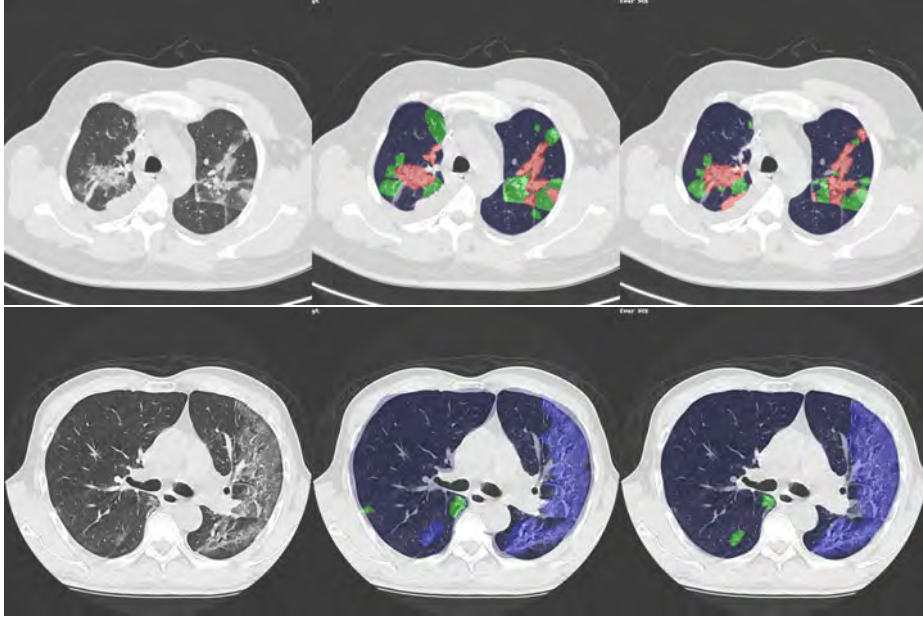
Fig. 3: Examples of U4 model's work. Red color is used for consolidation, blue for "crazy-paving", green for "ground-glass". On the left there is an original image, in the middle ground truth marking. On the right result of U4 BCE model's work.
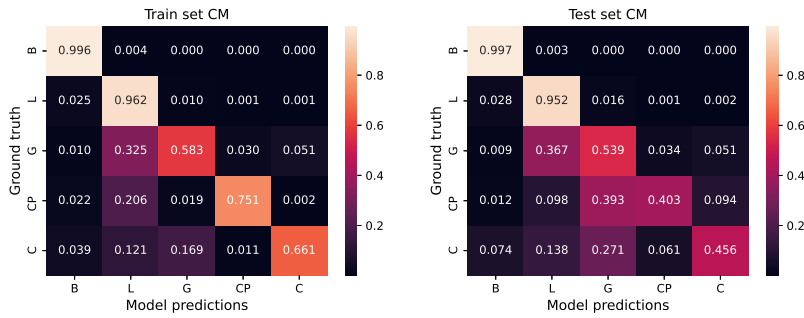

Fig. 4: Confusion matrices of U4 BCE model for train and test subsets. B refers to background, L to healthy lungs, G to "ground-glass", CP to "crazy-paving", and C to consolidation.

Table 5: Results for consolidation segmentation after dataset reduction.

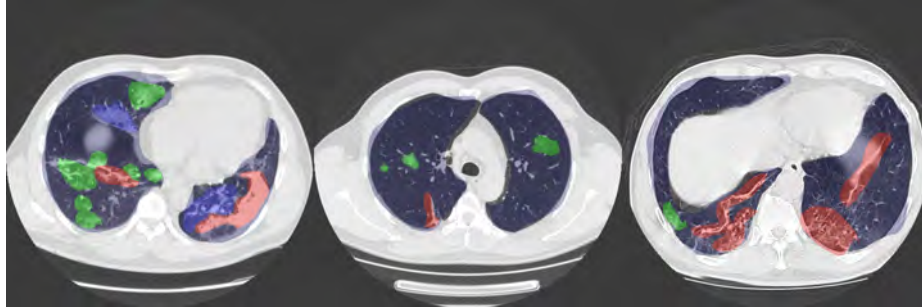| Experiment	mAP	IoU	Dice
C BCE	0.54077	0.2652	0.33648
C BCE (*)	**0.54998**	**0.30107**	**0.37059**

Fig. 5: Examples of inaccurate assessment.

consolidation class was lower than 0.2 in any of those models. Then new model C BCE (*) was trained on a new dataset.

As results show even rough cleaning of the dataset could improve the result of training.

Training of a "crazy-paving" model and a united model was not performed, because after dataset reduction too few examples of "crazy-paving" remained.

7   Conclusion

In this study we presented a solution for segmentation of lungs, lesions, pulmonary consolidation and "crazy-paving" with reasonable quality. Experiments showed that binary cross entropy loss works better for training models that could distinguish different types of pathologies, but gives slightly inferior results than the model trained with Dice loss for segmentation of general classes such as lungs and lesions. Then it was shown that usage of noisy data during training can decrease the quality of a model and discarding such data from the training subset can improve the quality of a model.

References

1. "Third Opinion Platform" Limited Liability Company, https://thirdopinion.ai/
2. He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., Xie, P.: Sample-efficient deep learning for covid-19 diagnosis based on ct scans. medRxiv (2020)
3. Huang, L., Han, R., Ai, T., Yu, P., Kang, H., Tao, Q., Xia, L.: Serial quantitative chest ct assessment of covid-19: Deep-learning approach. Radiology: Cardiothoracic Imaging **2**(2), e200075 (2020)
4. Jin, S., Wang, B., Xu, H., Luo, C., Wei, L., Zhao, W., Hou, X., Ma, W., Xu, Z., Zheng, Z., et al.: Ai-assisted ct imaging analysis for covid-19 screening: Building and deploying a medical ai system in four weeks. medRxiv (2020)
5. Liu, W., Zhang, C., Lin, G., Liu, F.: Crnet: Cross-reference networks for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4165–4173 (2020)
6. Qi, X., Jiang, Z., Yu, Q., Shao, C., Zhang, H., Yue, H., Ma, B., Wang, Y., Liu, C., Meng, X., et al.: Machine learning-based ct radiomics model for predicting hospital stay in patients with pneumonia associated with sars-cov-2 infection: A multicenter study. medRxiv (2020)

7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
8. Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shi, Y.: Lung infection quantification of covid-19 in ct images with deep learning. arXiv preprint arXiv:2003.04655 (2020)
9. Wang, L., Wong, A.: 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)
10. Zheng, C.: Time course of lung changes at chest ct during recovery from coronavirus disease 2019 (covid-19). Radiology **295**, 715–721 (2020)
11. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018)