COVID-CT-Dataset: A CT Scan Dataset about COVID-19

Jinyu Zhao   JIZ077@ENG.UCSD.EDU
UC San Diego

Yichen Zhang   YIZ037@ENG.UCSD.EDU
UC San Diego

Xuehai He   X5HE@ENG.UCSD.EDU
UC San Diego

Pengtao Xie   PENGTAOXIE2008@GMAIL.COM
UC San Diego, Petuum Inc

Abstract

CT scans are promising in providing accurate, fast, and cheap screening and testing of COVID-19. In this paper, we build a publicly available COVID-CT dataset, containing 275 CT scans that are positive for COVID-19, to foster the research and development of deep learning methods which predict whether a person is affected with COVID-19 by analyzing his/her CTs. We train a deep convolutional neural network on this dataset and achieve an F1 of 0.85 which is a promising performance but yet to be further improved. The data and code are available at https://github.com/UCSD-AI4H/COVID-CT

1. Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease that has affected 775,306 individuals all over the world and caused 37,083 deaths, as of Mar 30 in 2020. One major hurdle in controlling the spreading of this disease is the inefficiency and shortage of tests. The current tests are mostly based on reverse transcription polymerase chain reaction (RT-PCR). It takes 4-6 hours to obtain results, which is a long time compared with the rapid spreading rate of COVID-19. Besides inefficiency, RT-PCR test kits are in huge shortage.

This motivates us to study alternative testing manners, which are potentially faster, cheaper, and more available than RT-PCR, but are as accurate as RT-PCR. In particular, we are interested in CT scans. There have been several works (**???**) studying the effectiveness of CT scans in screening and testing COVID-19 and the results are promising. However, due to privacy concerns, the CT scans used in these works are not shared with the public. This greatly hinders the research and development of more advanced AI methods for more accurate testing of COVID-19 based on CT.

To address this issue, we build a COVID-CT dataset which contains 275 CT scans positive for COVID-19 and is open-sourced to the public, to foster the R&D of CT-based testing of COVID-19. From 760 medRxiv and bioRxiv preprints about COVID-19, we extract reported CT images and manually select those containing clinical findings of COVID-19 by reading the captions of these images. We trained a deep learning model on 183 COVID CTs and 146 non-COVID CTs to predict whether a CT image is positive for COVID-19.
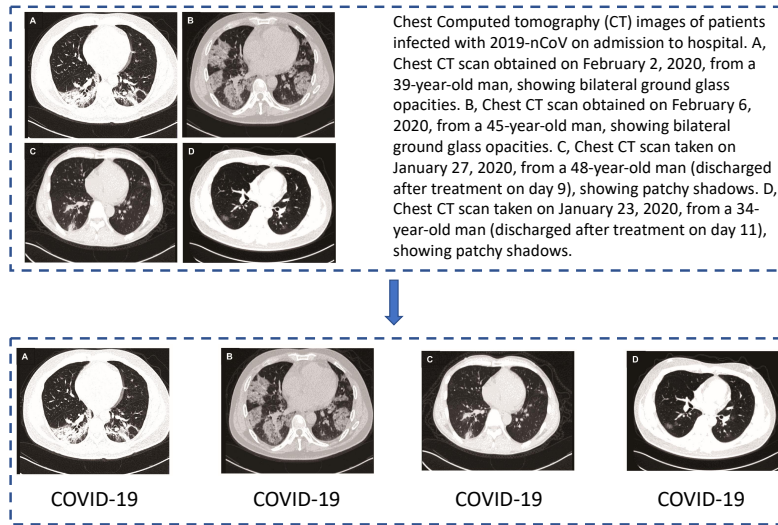
Chest Computed tomography (CT) images of patients infected with 2019-nCoV on admission to hospital. A, Chest CT scan obtained on February 2, 2020, from a 39-year-old man, showing bilateral ground glass opacities. B, Chest CT scan obtained on February 6, 2020, from a 45-year-old man, showing bilateral ground glass opacities. C, Chest CT scan taken on January 27, 2020, from a 48-year-old man (discharged after treatment on day 9), showing patchy shadows. D, Chest CT scan taken on January 23, 2020, from a 34-year-old man (discharged after treatment on day 11), showing patchy shadows.

COVID-19   COVID-19   COVID-19   COVID-19

Figure 1: For any figure that contain multiple CT scans as sub-figures, we manually split it into individual CTs.

Tested on 35 COVID CTs and 34 non-COVID CTs, our model achieves an F1 score of 0.85. The results demonstrate that CT scans are promising for screening and testing COVID-19, while more advanced methods are need to further improve the accuracy.

2. Dataset Creation

In this section, we describe how the COVID-CT dataset is built. We first collected 760 preprints about COVID-19 from medRxiv[1] and bioRxiv[2], posted from Jan 19th to Mar 25th. Many of these preprints report patient cases of COVID-19 and some of them show CT scans in the reports. CT scans are associated with captions describing the clinical findings in the CTs. We used PyMuPDF[3] to extract the low-level structure information of the PDF files of preprints and located all the embedded figures. The quality (including resolution, size, etc.) of figures are well-preserved. From the structure information, we also identified the captions associated with figures. Given these extracted figures and captions, we first manually select all CT scans. Then for each CT scan, we read the associated caption to judge whether it is positive for COVID-19. If not able to judge from the caption, we located the text analyzing this figure in the preprint to make a decision. For any figure that contain multiple CT scans as sub-figures, we manually split it into individual CTs, as shown in Figure **??**.

In the end, we obtain 275 CT scans labeled as being positive for COVID-19. These CT images have different sizes. The minimum, average, and maximum height are 153, 491, and

---

1. https://www.medrxiv.org/
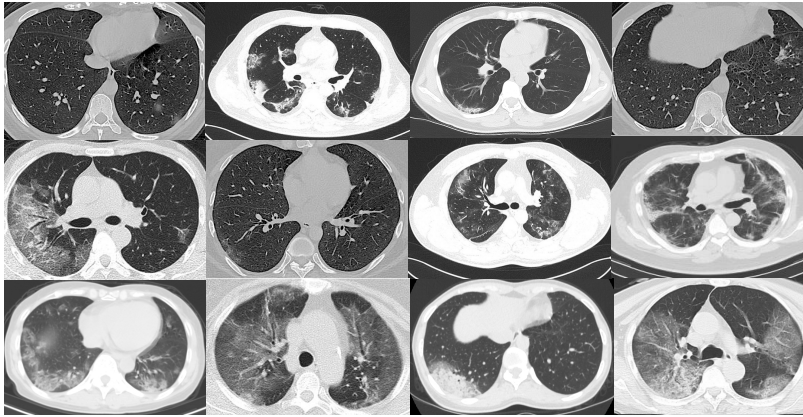2. https://www.biorxiv.org/
3. https://github.com/pymupdf/PyMuPDF

Figure 2: Examples of CT scans that are positive for COVID-19.

1853. The minimum, average, and maximum width are 124, 383, and 1485. These scans are from 143 patient cases. Figure ?? shows some examples of the COVID-19 CT scans.

3. Methods

We develop a baseline method on this dataset for the interested community to benchmark with. While our dataset is the largest publicly available CT dataset about COVID-19, it is still small. Training deep learning models on such a small dataset can easily lead to overfitting: the model would perform well on the training data, but generalizes badly on testing data. To address this problem, we adopt two approaches: transfer learning and data augmentation. Transfer learning aims to leverage a large collection of data from a relevant domain to help with the learning in the interested domain. Specifically, we use a large collection of chest X-ray images to pre-train a deep convolutional neural network, then fine-tune this pre-trained network on the COVID-CT dataset. Data augmentation aims to synthesize image-label pairs that are approximately correct, i.e., in most synthesized image-label pairs, the label is a correct annotation of the image.

3.1. Transfer Learning

To mitigate the deficiency of training data, we resort to transfer learning. Specifically, following (?) we use the ChestX-ray14 (?) dataset released by NIH to pretrain the DenseNet (?), then fine-tune the pre-trained DenseNet on the COVID-CT dataset.

3.2. Data Augmentation

Another way to mitigate data deficiency is data augmentation: from the limited training data, creating new image-label pairs and adding the synthesized pairs into the training set. We augment each training image by random affine transformation, random crop, and flip. The random affine transformation consists of translation and rotation (with degrees of 5, 15, 25).

4. Experiments

4.1. Experimental Settings

Table 1: Statistics of data split

	Non-COVID	COVID-19	Total
Train	146	183	329
Validation	15	57	73
Test	34	35	69

To train a binary classification model for predicting whether a CT image is COVID or non-COVID, we collect 195 CT scans that are negative for COVID. We split the dataset into a train, validation, and test set based on patients. Table **??** summarizes the number of COVID and non-COVID images in each set. All images are resized to 224-by-224. The hyperparameters are tuned on the validation set. The weight parameters in the networks were optimized using Adam (**?**) with a learning rate of 0.0001 with cosine scheduling and a mini-batch size of 4. We evaluate our method using five metrics: (1) Accuracy; (2) Precision; (3) Recall; (4) F1 score; (5) Area under ROC curve (AUC). For all metrics, the higher the better.

5. Results

Table 2: Performance (%) of the baseline method on the test set

Accuracy	Precision	Recall	F1	AUC
84.7	97.0	76.2	85.3	82.4

Table **??** shows the accuracy, precision, recall, F1, and AUC achieved by this baseline method. The precision is high but the recall is not satisfactory. More advanced methods are needed to improve recall.

6. Conclusions

We build a publicly available CT scan dataset about COVID-19, to foster the development of AI methods for using CT to screen and test COVID-19 patients. The dataset contains 275 CT scans that are positive for COVID-19. We train a deep learning model using this dataset and achieve an F1 score of 0.85. For the next step, we will continue to improve the method to achieve better accuracy.

References

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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*, page 200905, 2020.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *medRxiv*, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, Liang Yu, Yanfei Chen, Junwei Su, Guanjing Lang, et al. Deep learning system to screen coronavirus disease 2019 pneumonia. *arXiv preprint arXiv:2002.09334*, 2020.