Synthetic COVID-19 Chest X-ray Dataset for Computer-Aided Diagnosis

Hasib Zunair [1] **A. Ben Hamza** [1]

Abstract

We introduce a new dataset called Synthetic COVID-19 Chest X-ray Dataset [1] for training machine learning models. The dataset consists of 21,295 synthetic COVID-19 chest X-ray images to be used for computer-aided diagnosis. These images, generated via an unsupervised domain adaptation approach, are of high quality. We find that the synthetic images not only improve performance of various deep learning architectures when used as additional training data under heavy imbalance conditions (skew $\geqslant 90$), but also detect the target class with high confidence. We also find that comparable performance can also be achieved when trained only on synthetic images. Further, salient features of the synthetic COVID-19 images indicate that the distribution is significantly different from Non-COVID-19 classes, enabling a proper decision boundary. We hope the availability of such high fidelity chest X-ray images of COVID-19 will encourage advances in the development of diagnostic and/or management tools.

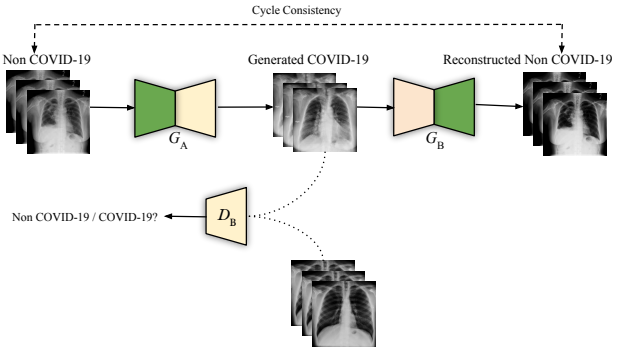


Figure 1. Illustration of the data generation process based on unpaired image-to-image translation. CXR images are translated from Non-COVID-19 (i.e. Normal or Pneumonia) to COVID-19 and then back to Non-COVID-19 via cycle-consistency.

1. Introduction

Recent studies have shown that chest radiography images such as chest X-rays (CXR), performed on patients with COVID-19 when they arrive at the emergency room, can help doctors determine who is at higher risk of severe illness and intubation (Ai et al., 2020; Huang et al., 2020). Automatic interpretation of chest radiography images such as CXR using computational approaches not only helps healthcare organizations save time and money, but also provides superior patient care and more importantly it can save lives (Ng et al., 2020).

Several computational approaches for the detection of COVID-19 cases from chest radiography images have been recently proposed, including tailored convolutional neural network (CNN) architectures (Karim et al., 2020; Wang et al., 2020) and transfer learning based methods (Kassani et al., 2020; Narin et al., 2021; Li et al., 2020; Farooq & Hafeez, 2020). While promising, the predictive performance of these deep learning based approaches depends heavily on the availability of large amounts of curated and annotated data. Over the past few years, there have been several efforts to build large-scale annotated datasets for CXRs and make them publicly available to the global research community (Johnson et al., 2019; Wang et al., 2017; Bustos et al., 2020). At the time of writing, there exists, however, only one annotated COVID-19 X-ray Image Data Collection (Cohen et al., 2020), which is a curated collection of CXR images of patients who are positive or suspected of COVID-19 or other viral and bacterial pneumonia. This is largely attributed to the rare nature of the radiological finding, legal, privacy, technical, and data-ownership challenges. While the COVID-19 Image Data Collection contains positive examples of COVID-19, the negative examples were acquired from publicly available sources (Wang et al., 2017) and merged together for computational analysis. This fusion of multiple datasets results in predominantly negative examples with only a small percentage of positive ones (i.e COVID-19), giving rise to a class imbalance problem (Johnson et al., 2019; Wang et al., 2017; Bustos et al., 2020).

[1]Concordia University, Montreal, QC, Canada. Correspondence to: Hasib Zunair <hasibzunair@gmail.com>.

Accepted at the ICML 2021 Workshop on Computational Biology (WCB). Copyright 2021 by the author(s).

[1]GitHub: Synthetic COVID-19 CXR Dataset

2. Methods

Identifying COVID-19 in a CXR image can be regarded as an image classification problem. We consider two binary classification tasks, namely Normal vs. COVID-19 and Pneumonia vs. COVID-19, for which we want to accurately identify COVID-19. For training the classification and the generative models, we use COVID-19 samples from the COVID-19 Image Data Collection (Cohen et al., 2020), and Normal and Pneumonia samples from the RSNA Pneumonia Detection Challenge (Wang et al., 2017).

In order to address the class imbalance problem, we recently proposed an unsupervised domain adaptation algorithm to synthesize under-represented class samples (i.e COVID-19 CXR images) from the over-represented ones (i.e Normal or Pneumonia CXR images) using unpaired image-to-image translation (Zunair & Hamza, 2021; 2020). The key idea is to train a cycle-consistent generative model (Zhu et al., 2017) on unpaired samples of two domains (Non-COVID-19 and COVID-19 in our case) with the goal of learning mapping functions between them. We train two translation models, which learn the mapping from Normal to COVID-19 and Pneumonia to COVID-19. After training, given a Non-COVID-19 CXR image (i.e. Normal or Pneumonia) as input, the generative model translates such that the image has representative features of COVID-19. An illustration of the data generation process in shown in Figure 1.

We generate 16,537 and 4,758 COVID-19 CXR images for Normal vs. COVID-19 (denoted $\mathcal{G}_{NC}$) and Pneumonia vs. COVID-19 (denoted $\mathcal{G}_{PC}$) tasks, respectively.

3. Results

We report results in Figure 2 for Normal vs. COVID-19 and Pneumonia vs. COVID-19 tasks when using $\mathcal{G}_{NC}$, instead of $\mathcal{G}_{PC}$, as additional training data across various deep learning architectures. It is evident that for all architectures, adding the synthetic data significantly improves COVID-19 detection performance. For Normal vs. COVID-19, the performance is much better when adding $\mathcal{G}_{NC}$ instead of $\mathcal{G}_{PC}$. We hypothesize that this is due largely to the fact that the number of images in $\mathcal{G}_{NC}$ is much larger.

In Figure 3, we display the two-dimensional Uniform Manifold Approximation and Projection (UMAP) embeddings of the features with the objective of visualizing the difference between the original and synthetic data. Figure 3(a) shows that the original examples exhibit low interclass variation and consist of outliers. In Figures 3(b) and 3(c), the synthetic samples are in a different distribution in the feature space. While the UMAP embeddings may not be interpreted as a justification that the synthetic examples actually consist of COVID-19 symptoms from a clinical perspective, it is, however, important to note that the distribution of the synthetic images is significantly different than that of normal images; thereby enabling a proper decision boundary.

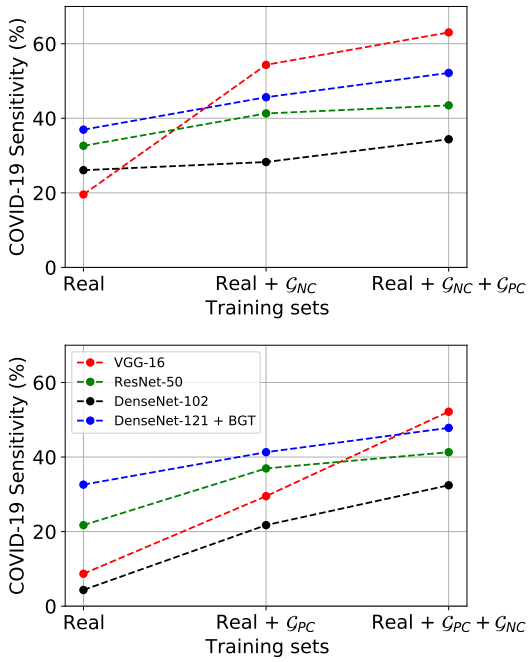
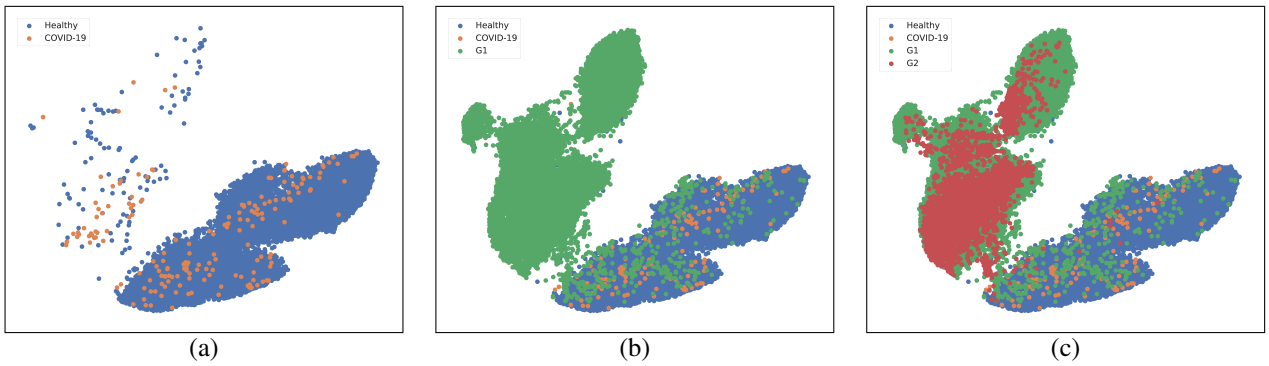


Figure 2. COVID-19 detection performance results on Normal vs. COVID-19 (top) and Pneumonia vs. COVID-19 (buttom) test sets when trained on real data, and on combined real and synthetic data. For both tasks, synthetic data improves detection performance of the different classification models.

To visually explain the decisions made by the model in the sense that why an X-ray image is classified as COVID/Non-COVID, we use the gradient-weighted class activation map (Grad-CAM) to generate the saliency maps that highlight the most influential features affecting the predictions. Since the convolutional feature maps retain spatial information and that each pixel of the feature map indicates whether the corresponding visual pattern exists in its receptive field, the output from the last convolutional layer of the deep neural network shows the discriminative region in an image. To distinguish between the predicted COVID-19 and Non-COVID-19 images, we visualize the saliency maps for images that are correctly classified as COVID-19 and Non-COVID-19 (normal) by the proposed model. As shown in Figure 4, the class activation maps for Non-COVID-19 (normal) demonstrate high activations for regions around the lungs, suggesting that there are no prediction features indicating that the disease is present. For most of the images that are correctly classified as COVID-19, the highlighted regions are within the lungs. Notice that in some cases, the model only highlights a specific part of the lung (e.g. left or right), which shows that COVID-19 features are present only on one side.



(a) (b) (c)

Figure 3. Two-dimensional UMAP embeddings: (a) Normal vs. COVID-19; (b) Normal vs. COVID-19 + $\mathcal{G}_{NC}$; (c) Normal vs. COVID-19 + $\mathcal{G}_{NC}$ + $\mathcal{G}_{PC}$; Here, G1 and G2 denote $\mathcal{G}_{NC}$ and $\mathcal{G}_{PC}$, respectively.

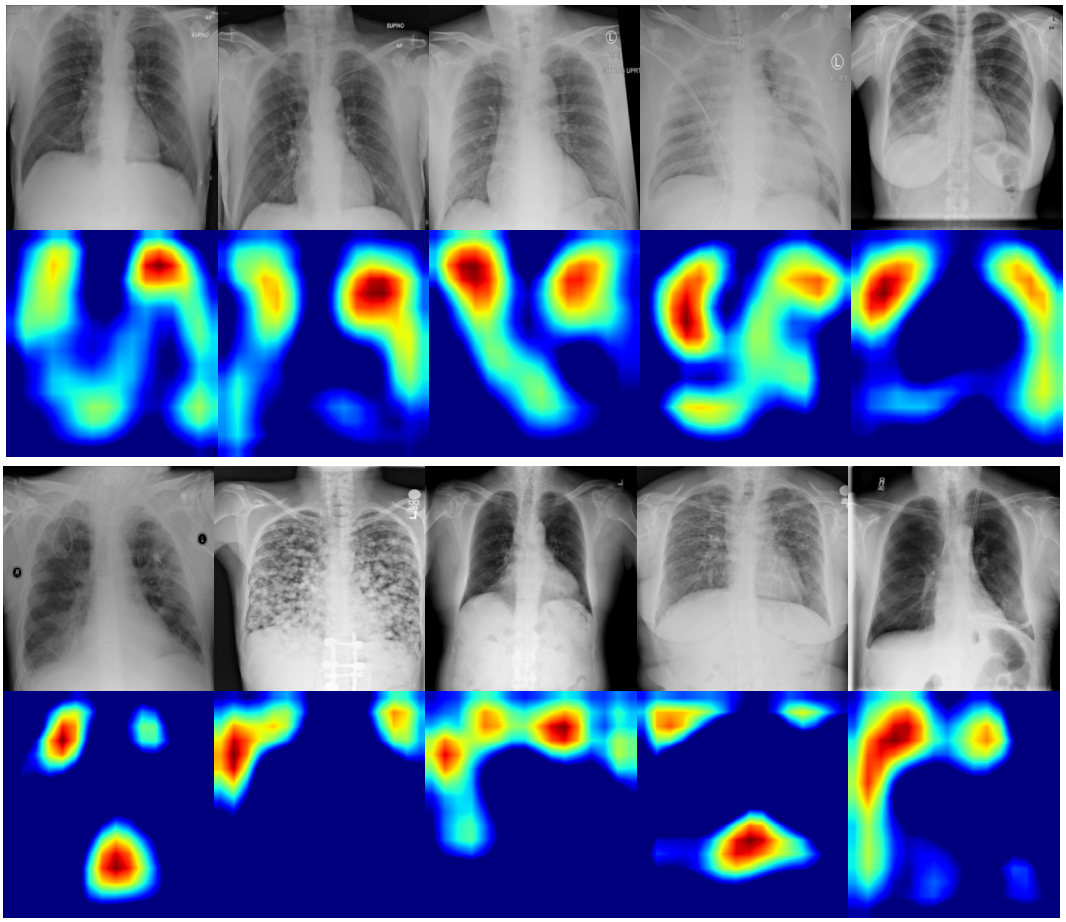


Figure 4. Saliency maps for the correctly classified COVID-19 (top two rows) and Non-COVID-19 (bottom two rows) images by the proposed model. Notice that for images that are classified as COVID-19, our model highlights the areas within the lungs, whereas for Non-COVID-19 images, the most important regions are around the lungs.

4. Conclusion

We release a publicly available dataset consisting of 21,295 synthetic COVID-19 CXR images to be used for training machine learning models in computer-aided diagnosis. We find that using these images can alleviate heavy class imbalance problems across multiple deep learning architectures for COVID-19 detection. Salient features also suggest that the distribution of synthetic images are different from other classes, and hence enable a proper decision boundary.

References

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*, 2020.

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. COVID-19 image data collection: Prospective predictions are the future. *Journal of Machine Learning for Biomedical Imaging*, 2020.

Farooq, M. and Hafeez, A. COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs. *arXiv preprint arXiv:2003.14395*, 2020.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, 2020.

Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., ying Deng, C., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Karim, M., Döhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M., Beyan, O., et al. DeepCOVIDExplainer: Explainable COVID-19 predictions based on chest X-ray images. *arXiv preprint arXiv:2004.04582*, 2020.

Kassani, S. H., Kassasni, P. H., Wesolowski, M. J., Schneider, K. A., and Deters, R. Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: A machine learning-based approach. *arXiv preprint arXiv:2004.10641*, 2020.

Li, X., Li, C., and Zhu, D. COVID-MobileXpert: On-device COVID-19 screening using snapshots of chest X-Ray. In *Proc. IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1063–1067, 2020.

Narin, A., Kaya, C., and Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, pp. 1–14, 2021.

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. Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic Imaging*, 2(1), 2020.

Wang, L., Lin, Z. Q., and Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. *Scientific Reports*, 10(1):1–12, 2020.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

Zunair, H. and Hamza, A. B. Melanoma detection using adversarial training and deep transfer learning. *Physics in Medicine & Biology*, 65, 2020.

Zunair, H. and Hamza, A. B. Synthesis of covid-19 chest x-rays using unpaired image-to-image translation. *Social Network Analysis and Mining*, 11(1):1–12, 2021.