CONDITIONAL SYNTHETIC DATA GENERATION FOR ROBUST MACHINE LEARNING APPLICATIONS WITH LIMITED PANDEMIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Background: At the onset of a pandemic, such as COVID-19, data with proper labeling/attributes corresponding to the new disease might be unavailable or sparse. Machine Learning (ML) models trained with the available data, which is limited in quantity and poor in diversity, will often be biased and inaccurate. At the same time, it is imperative that ML algorithms to fight pandemics must have good performance and be developed in a time-sensitive manner. To tackle the challenges of limited data, we propose generating conditional synthetic data, to be used alongside real data for developing robust ML models. **Methods:** We present a hybrid model consisting of a conditional generative flow and a classifier for conditional synthetic generation. The classifier decouples the feature representation corresponding to the condition, which is fed to the flow to extract the local noise. We generate synthetic data by manipulating the local noise with fixed conditional feature representation. **Results:** We performed conditional synthetic generation for chest computed tomography (CT) images corresponding to normal, COVID-19 and pneumonia afflicted patients. We show that our method significantly outperforms existing models both on qualitative and quantitative performance. As an example of downstream use of synthetic data, we show improvement in COVID-19 detection from CT scans with conditional synthetic data augmentation.

1 INTRODUCTION

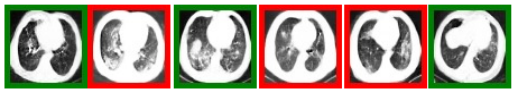

Figure 1: Synthetic images generated by our proposed model, with Non-COVID (normal and pneumonia cases, images with green border)/ COVID (images with red border) as the condition.

The COVID-19 pandemic has created a public health crisis and continues to have devastating impact on lives and healthcare systems worldwide. In the fight against this pandemic, a number of algorithms involving state-of-the-art machine learning techniques have been proposed. Data-based approaches have been used in a number of important tasks such as detection, mitigation, transmission modeling, decision on lockdown, reopening and related restrictions etc. For example, computer vision-based detection of COVID-19 from chest computed tomography (CT) images has been proposed as a supportive screening tool for COVID-19 (Gunraj et al., 2020), along with the primary diagnostic test of transcription polymerase chain reaction (RT-PCR). This is beneficial since obtaining definitive RT-PCR test results may take a lot of time in critical situations.

The application of machine learning algorithms in healthcare depends upon ample availability of quality data along with their attributes/labels. At the beginning of a pandemic, labeled data corresponding to the disease might be unavailable or sparse. Sparse data often has limited variation in several important factors relevant to disease detection such as age, underlying medical conditions etc. Class imbalance is another issue faced by machine learning algorithms when pandemic-disease related data is limited. For example, at the onset of COVID-19, the amount of CT scan images corresponding to COVID-19 were much less than those corresponding to existing lung diseases (e.g. pneumonia). ML models fed with such class-imbalanced data could be biased and thus provide

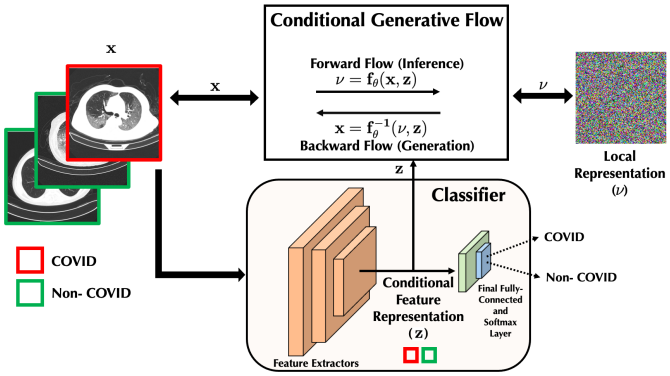

Figure 2: Illustration of the proposed conditional synthetic generation. (Best viewed in color)

inaccurate results. Concurrently, after a new disease has been discovered, the healthcare ML tools must rapidly adapt to the new disease in order to assist medical professionals diagnose and treat susceptible or affected individuals as quickly as possible.

In this paper, we present a novel conditional synthetic data-generation method to augment the available pandemic data of interest. At the onset of a pandemic, when the availability of labelled data is limited, our proposed model learns the distribution of available limited data and then generates conditional synthetic data that are added to the existing data in order to improve the performance of machine learning algorithms. This can enable healthcare ML tools to rapidly adapt to the pandemic.

We apply this method to generate conditional CT scan images corresponding to COVID cases, and conduct qualitative and quantitative tests to ensure that our model generates high-fidelity samples and is able to preserve the features corresponding to the condition in synthetic samples. As a downstream use of conditional synthetic data, we improve the performance of COVID-19 detectors based on CT scan data via synthetic data augmentation. Our results show that the proposed model is able to generate synthetic data that mimic the real data, and the generated image samples can indeed be augmented with existing data to improve COVID-19 detection efficiency.

2 METHODOLOGY

We present a hybrid model consisting of a conditional generative flow and a classifier for conditional synthetic generation. The following sections describe the operation of both the components.

2.1 COVID AND NON-COVID CLASSIFIER

Our model is characterized by the efficient decoupling of feature representations corresponding to the condition and the local noise. We first train the classifier to classify the input sample (which in our case are CT Scans) and associated labels as COVID and Non-COVID. By virtue of the training process, the classifier learns to discard the local information and preserve the features necessary for classification (conditional information) towards the downstream layers. Once the classifier is trained, we freeze its parameters, and use it to extract the conditional feature representation z (as a vector without spatial characteristics) for input image x while training the conditional generative flow. We extract z from close to the penultimate layer of the classifier so as to ensure the information represents the conditions (COVID/Non-COVID). The dimension of z is chosen to be significantly lower than the input data, i.e. $\dim(z) << \dim(\text{input data})$.

2.2 CONDITIONAL GENERATIVE FLOW

During the inference phase for the flow model, the conditional feature representation z is fed to the conditional generative flow, which transforms x to its local representation ν, i.e. $f_\theta(x, z) = \nu \sim \mathcal{N}(0, I)$, with the same dimension as x by the inherent design of flow models. For more details on how z is incorporated as a conditional input to the flow model, please refer to Ma et al. (2021). Since

	Quantitative				Qualitative
	Accuracy	**F1**	**Precision**	**Recall**	**FID**
Ma et al. (2021)	0.8419	0.7115	0.7771	0.6830	0.2504
ACGAN	0.7938	0.7458	0.7293	0.8241	0.0986
CAGlow	0.8681	0.7909	0.8017	0.7816	0.0483
Ours	**0.9574**	**0.9318**	**0.9537**	**0.9134**	**0.0077**
Real Data	0.9646	0.9453	0.9495	0.9412	–

Table 1: Conditional synthetic data generation results (for FID, the smaller the better)

% COVID	**Accuracy**	**F1**	**Precision**	**Recall**
Real data ($\sim$ 20%)	0.9646	0.9453	0.9495	0.9412
30%	**0.9719**	**0.9567**	**0.9585**	0.9549
40%	0.9659	0.9474	0.9504	0.9445
50%	0.9717	0.9566	0.9568	**0.9564**

Table 2: Results for classification models trained using augmented training data with varying % of COVID +ve samples

flow models are bijective mappings, the exact x can be reconstructed by the inverse flow with z and ν as inputs. During the generation phase, for an input sample x, keeping the conditional feature representation the same (z), we sample a new local representation $\tilde{\nu}$, to generate a synthetic sample $\tilde{x} = f_\theta^{-1}(\tilde{\nu}, z)$, with the same conditional features (Same COVID/Non-COVID features) as x , but a different local representation. An illustration of the proposed model is provided in Fig. 2.

3 Experiments

3.1 Data Collection and Pre-processing

We conduct experiments on chest CT scan data based on the COVIDx CT-1 dataset (Gunraj et al., 2020). This dataset consists of 45,758 images corresponding to healthy cases, 36,856 images corresponding to common pneumonia cases, and 21,395 images corresponding to COVID-19 cases.

Pre-processing: We combine the images in the Normal and Pneumonia classes into a single "Non-COVID"class. We use the train, validation, and test splits defined by the official annotations files. We crop the images as per the given bounding box for lung region and resize them to 64×64.

3.2 Testing Procedure

We performed both quantitative and qualitative testing for conditional synthetic data generation by our model. A test set is held out from the real dataset to be used for quantitative testing. We then compare the classification performance (COVID/Non-COVID) on this test set for a classifier trained on real data vs a classifier trained on the generated synthetic data. Since the datasets are imbalanced, we report the precision, recall and macro-F1 score along with the accuracy. For more information on the metrics, please refer to Mishra (2018). To evaluate the quality of generated samples, we report the Fréchet Inception Distance (FID) (Heusel et al., 2018) for the synthetic samples. For FID calculation, we use the embeddings from our classifier trained using real data, in place of the official inception network (Szegedy et al., 2014), since it is not trained on the medical imaging data.

3.3 Conditional Synthetic Data Generation Results

The classification results for a classifier trained on the real data vs a classifier trained on purely conditional synthetic data is given in Table 1. The table also includes the FID scores for generated synthetic data. Across the existing methods for conditional synthetic generation, the classifier trained with synthetic data from our method has the closest accuracy and F1 score to that of the classifier trained on real data. This shows the capability of our method to generate synthetic samples with a distribution that closely matches the real conditional data distribution. The FID scores for the synthetic data generated using our model are the lowest among all models, demonstrating that the efficient quality of generated samples.

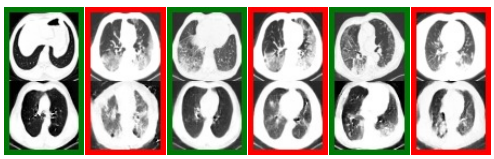

Figure 3: Original and generated synthetic samples for CT scans. Top row: original samples, Bottom row: synthetic sample corresponding to the image top row. Image pairs with a red border: COVID samples, and a green border: Non-COVID samples.

The original samples along with the synthetic samples generated by preserving original conditional feature representation and a different local noise for CT scans are shown in Fig. 3. The characteristic features for COVID CT scan samples, i.e., ground-glass opacity are well preserved in the synthetic samples. At the same time, the non-conditional local features, e.g. axial plane position for CT scans are considered as local noise. Since original samples for normal and pneumonia cases are merged together to form a single Non-COVID class, sometimes the corresponding synthetic image for a normal sample is a sample with pneumonia characteristics and vice-versa. This occurs since the conditional model learns to treat them as local information. The ability to decouple the feature representations for given conditions from other information in the data, as exhibited by our model, should be considered the strength of an effective conditional generative model.

3.4 Robust Detection of COVID-19 via Synthetic Data Augmentation

Generated synthetic data can be utilized in a number of downstream tasks, including robust detection of COVID-19 via synthetic data augmentation. The training data is inherently highly class-imbalanced, with limited samples of COVID and abundant samples for pneumonia and normal cases. To design a robust COVID-19 detection mechanism under such class imbalance scenario, we augment the training data with synthetic COVID samples generated using the proposed model to increase the % of COVID samples to 30%,40%,50%. The classification metrics for classifiers trained on the augmented training data are given in Table 2. Examining the classification results, the classifiers trained on augmented training data have better performance as compared to classifiers trained only on limited real training data for all augmentation levels. Note that even slight improvement in the recall score translates to numerous samples classified correctly (e.g. 1% improvement in recall for CT scan corresponds to 200 more correctly classified samples), leading to better diagnosis.

4 Related Work

In the field of healthcare, synthetic data generation has been proposed to expand the scope and amount of the existing training data, often to improve the robustness of ML models. Ghorbani et al. (2019) propose a generative adversarial network (GAN)-based synthetic data generation to improve the diversity and the amount of skin lesion images. Kohlberger et al. (2019) synthesize pathology images for cancer. Waheed et al. (2020) propose a ACGAN-based (Odena et al., 2017) generator for conditional synthetic chest X-ray data generation and augmentation for robust COVID-19 detection.

In the area of conditional generation, a hybrid flow and a GAN-based model have been proposed in CAGlow (Liu et al., 2019). In general, GAN-based methods are known to be hard to train (Salimans et al., 2016) and do not provide a latent embedding suitable for feature manipulations (Kingma & Dhariwal, 2018). In contrast, we proposed a conditional generation method with efficient decoupling of the conditional information and local noise over an embedding space, along with a flow based generator, which in recently have proved efficient in synthetic data generation (Ho et al., 2019).

Decoupling of global and local representation for synthetic generation has been proposed in Ma et al. (2021), where the global information is decoupled using a Variational AutoEncoder (VAE) (Kingma & Welling, 2014) network. For conditional synthetic generation, it is necessary that the feature representations salient to the given conditions (COVID vs Non-COVID) are decoupled from local noise, which is not guaranteed while extracting the same using a VAE. By employing a classifier network for the same, we ensure the relevant conditional information is not lost into the local noise.

5 Conclusions and Future Work

We presented a novel conditional synthetic generative model, aimed at multiplying the samples of interest at the onset of a pandemic. We conducted extensive experiments on chest CT scan dataset to show the efficacy of the proposed model, and improvements in COVID-19 detection performance achieved via synthetic data augmentation. With appropriate changes in the networks and training mechanism, our method can be generalized for synthetic generation of other kinds of data, e.g. X-rays, natural language and time-series. We believe our proposed conditional synthetic data generation work will enable new avenues of research into synthetic realization of medical data and eventually robust models in healthcare AI, essential in the fight against future pandemics.

REFERENCES

Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. Dermgan: Synthetic generation of clinical skin images with pathology, 2019.

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, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design, 2019.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

Timo Kohlberger, Yun Liu, Melissa Moran, Po-Hsuan Cameron Chen, Trissia Brown, Jason D Hipp, Craig H Mermel, and Martin C Stumpe. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. *Journal of pathology informatics*, 10, 2019.

Rui Liu, Yu Liu, Xinyu Gong, Xiaogang Wang, and Hongsheng Li. Conditional adversarial generative flow for controllable image synthesis, 2019.

Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard H Hovy. Decoupling global and local representations via invertible generative flows. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iWLByfvUhN.

Aditya Mishra. Metrics to evaluate your machine learning algorithm. 2018. URL https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans, 2017.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access*, 8: 91916–91923, 2020. doi: 10.1109/ACCESS.2020.2994762.