Automated Detection of COVID-19 using Convolutional Neural Networks and Generative Adversarial Networks

*This report was written as part of a Masters of Science in Artificial Intelligence and Big Data Analytics

Ultan Kearns
Department of Computing
ATU Letterkenny

I. Abstract

This report was written as a requirement for a Masters of Science in Artificial Intelligence and Big Data Analytics at ATU Letterkenny and aims to summarize the results of the research into synthetic data generation to train CNNs for the purpose of automated detection of COVID-19. The limited amount of data available for training CNNs to recognize COVID-positive and COVID-negative patients has led to a number of researchers training their models on mislabelled or imbalanced datasets, the results of this research have shown that there is some promise to synthetically augmenting data through the use of generative deep-learning. Throughout the research we used a variety of different methods to generate the new synthetic data and to balance the dataset by augmenting minority classes to bring them in balance with the majority classes. The results of this research have shown that the use of synthetic augmentation can improve the accuracy of some CNN models and although the CNN models were unable to surpass the currently achieved validation set accuracy of the COVID models mentioned in the literature review(the majority of these models attained a validation set accuracy of $> 98\%$)[1][2] it is worth nothing that the top model(Extensive CNN CT EfficientNetV2S) trained achieved a test set accuracy of 96% when trained on a larger dataset which contained 4,655 more images than the model mentioned in the literature review. It is also worth noting that this research was conducted using a test set whereas the model mentioned in the literature review only used a validation set which shows that the researchers may have overfit the validation set.

II. Introduction

The crux of the research conducted centred around the question ”can generative deep-learning be used to improve the results of currently existing models used to automate the diagnosis of COVID-19?” When beginning the study we researched various different architectures of GAN and various architectures of CNNs. The main GAN architecture used in this study is the Deep Convolutional GAN, this was the GAN which had the most success in generating synthetic images. In addition to the DCGAN Variational Autoencoders(VAEs) were also used but had limited success when augmenting the datasets. Transfer learning was also used with a number of different pre-trained models all trained using the ImageNet dataset being used to automate COVID-19 diagnosis and many were able to outperform the baseline models. The research goal was achieved upon conclusion of this thesis with many of the models trained on the augmented datasets outperforming the models trained on the non-augmented datasets.
The research conducted used a number of datasets to train both the CNNs and the GANs. Three datasets were used to train the GAN models these were the COVID-19 Radiography Dataset, the Extensive COVID-19 dataset, and finally the COVID-19 X-Ray dataset. There was also a fourth dataset used to train CNN models the COVID-19 Chest X-Ray dataset. However this dataset was not suitable to train GAN models as there are 11 classes and some classes are extremely underrepresented so in an effort to conserve resources the decision was made not to use this dataset when training GANs. The CNN models for this dataset were still included as the effect of transfer-learning was still interesting to see when using a severely limited dataset.

III. Literature Review Findings

A number of papers were reviewed during the course of this research to determine current paradigms in generative deep-learning and computer vision. This section began with researching and analysing currently existing models which were used in the automated detection of COVID-19 in patients. In a paper by Mahmoudi, Benamour et al[2] researchers investigated deep learning approaches to aid in the creation of an automated diagnostic tool using computed topography scans. The researchers found that segmenting the infected region of the patients scan and using a technique called Contrast Limited Adaptive Histogram Equalization was beneficial to creating a homogenous dataset and aiding in the improvement of the model's performance. The researchers removed black slices from the images so that only the region of interest was shown and used a U-net architecture for more timely and accurate segmentation of images. Four-fold cross-validation was also used in the training of this model. The model achieved an accuracy of 98% when diagnosing patients. Despite the high accuracy the model suffered from lack of data as only 20 CT Scans were used to train and validate the model(40 images in total). The model was also deprived of a test set which may have inflated the accuracy as it is highly likely the model was overfitting to the validation set. Due to the lack of data and the high-cost associated with data-collection the researchers were unable to collect more data for a test-set to provide a less biased and more objective result.

There is also a high potential for bias in this study as the limited amount of data and possibly mislabelled or misclassified data may have had an adverse effect on analyzing the results. This links back to a term ”Frankenstein Datasets” these are datasets which are mislabelled, suffer from lack of data-quality, biased, and spliced together from different sources. Frankenstein datasets were used initially in the early-stages of the pandemic as many researchers and companies rushed to find a solution to training a high quality automated COVID-19 diagnostic tool to alleviate the heavy burden placed upon medical staff at the time.

The issue with limited data and imbalance between classes within the datasets is shown in the next study analyzed also and seems to be a recurring theme when analyzing currently existing automated COVID-19 diagnostic tools. In a paper by Islam, Islam, and Asraf[1] a new method was used to automate the diagnosis of COVID-19 called LSTM(Long Short Term Memory). The model trained as part of this research used a much larger dataset than the study analyzed previously[2], the dataset used contained 4,575 X-ray images in total with 1,525 images being COVID-19 Positive. The model achieved an area under curve accuracy of 99.4%, an accuracy of 99%, a specificity of 99.2%, and a sensitivity of 99.3%. The researchers suggest that the study may have been improved if more data was available to them, the lack of data is clearly visible from the size of the dataset and the imbalance between classes. The model does appear to improve upon the initial model mentioned in the previous study[2] as it is trained on a far larger dataset and would have an increased ability to generalize. As shown when comparing the accuracy of this model with the one mentioned previously this model has a much higher accuracy.

The model trained as part of this study appeared to greatly benefit from the use of an LSTM architecture. The use of which allows the neural network to remember and carry previously learned information to layers deeper in the network, this greatly helps with the problems of vanishing and exploding gradients[3].

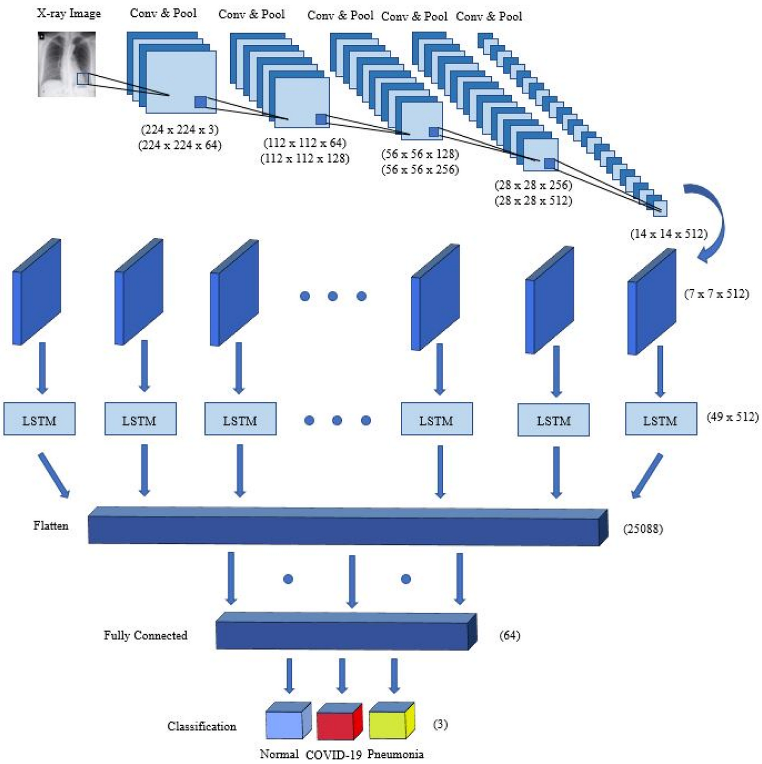

Fig. 1. Overview of a Typical LSTM Network[1]

Figure 1 shows the architecture of an LSTM Network as shown above the LSTM is comprised of four main gates an input gate, ”a forget gate”, an update gate, and finally an output gate. The gates essentially determine which data the model should remember and which data isn’t useful for training and can be forgotten. The models trained achieved a much better result than the model in the previous study and when considering the much larger size of the dataset the model trained as part of this study would have a much higher accuracy when it came to generalizing and classifying new data.

In the literature review we also address the challenges and limitations of using Artificial Intelligence in developing automated diagnosis systems for COVID-19 classification. In a paper by Huang, Yang and others[4], researchers discuss the current challenges and limitations of developing an Artificial Intelligence model to assist medical professionals in the identification and diagnosis of COVID-19. The challenges include: lack of data, lack of data quality, and the use of poorly merged data sets which we discussed earlier(”Frankenstein datasets”). There are more challenges other than those previously discussed such as finding patients who are COVID-19 positive and asymptomatic(most will not seek medical diagnosis as they don’t present symptoms), the labelling of data is also an issue as there is a risk of mislabelled data by clinicians given moderate signs of COVID-19 being hard to definitively classify as COVID positive or negative. The risk of false positives and false negatives both of which present issues such as unnecessary quarantine and further spread of infection.

To mitigate these challenges the researchers of this paper suggest that when developing an AI diagnostic system that developers should combine chest imaging, exposure history, and laboratory tests when training / testing the model. The issue with this approach is laws concerning data-collection and ethical questions regarding the right to the patients privacy. There is also a high cost associated with data-collection and to collect enough data may take quite a long time.

The issues explored in the literature review regarding current constraints to developing automated diagnostic models for COVID-19 detection lead into the next section Research into Data Augmentation and Convolutional Neural Networks. In this section we aim to analyze the use of augmenting datasets with synthetic data created by GANs to see if there have been positive results across different problem domains. The section opens with an explanation of synthetic data and discussing that there are numerous techniques and methodologies for using data augmentation, for the purpose of this study we will be using GANs to generate new data from currently existing data. Then we move on to discuss some of the advantages of using data augmentation for training the models, the main advantages are: larger training set to train models on, reduces overfitting, helps to prevent underfitting and improve the accuracy of the model, and reduces the cost/time associated with gathering new data, and increases the models ability to generalize. We also discuss the limitations of using GANs the key limitations are: inability to reduce bias(if bias is present in the dataset the synthetic set will also contain bias), hard to generate discrete data(this had little effect on our research), and the data generated will need to be evaluated(some of the synthetic data generated in this study appeared malformed, unfortunately due to the time associated with pruning such data and the computational resource limitations such malformed images are present in the augmented sets).

After discussing the advantages and limitations of using synthetically augmented data we then moved on to discuss a paper by Tanaka and Aranha[5]. In this paper the researchers discuss two algorithms SMOTE(Synthetic Minority Over-sampling Technique) [6]and ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning)[7]. Both techniques are used to oversample the minority class within datasets to balance the dataset.

We then discuss how each algorithm works. SMOTE works by creating artificial data which takes into account the data's position, a random point in the least represented class in the dataset is selected and SMOTE identifies members of the same class within the data by using the k-nearest neighbour algorithm which is a form of unsupervised learning. SMOTE then generates an entirely new point in the vector for each pair which is situated between the two pieces of data, the new point is then positioned at a random percentage away from the initial point chosen[5]. ADASYN works in a similar way to SMOTE and was originally based on SMOTE. Both function in much the same way but the key difference lies in ADASYN adding a random small bias value to the points, breaking linear correlation to their parents. The bias that ADASYN adds helps to increase the amount of variance within the synthetic data. We then evaluate the performance of each of these algorithms. When testing both SMOTE and ADASYN the researchers found that both underperformed in accuracy on the imbalanced dataset compared to the original model but had a higher recall. When the datasets were balanced it was found that SMOTE and ADASYN performed much better. The models also achieved an increase in accuracy and performance when trained using traditional GANs. By reviewing this study it showed that the use of GANs were able to improve accuracy of certain classification models across a number of problem domains. The researchers used 3 datasets when training the models and synthetically augmenting the datasets. These datasets were

- Pima Indians Diabetes data Database
- Breast Cancer Wisconsin Data Set (Diagnostic)
- Credit Card Fraud Detection

Given that the researchers had shown success in using GANs across these vastly different problem domains it showed that there may be some value into researching syntehtic data augmentation when it came to alleviating data-shortages in COVID-19 diagnosis.

The next study analyzed as part of researching the current use of GANs to augment imbalanced datasets was from a paper by Wang and Xiao[8].

IV. Design & Implementation

V. Results

VI. Conclusion

References

[1] I. Islam and Asraf, “A Combined Deep CNN-LSTM Network for the Detection of Novel Coronavirus (COVID-19) using X-ray Images,” *Informatics in Medicine Unlocked*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820305621.

[2] B. Mahmoudi *et al.*, “A Deep Learning-Based Diagnosis System for COVID-19 Detection and Pneumonia Screening Using CT Imaging,” *Applied Sciences*, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/10/4825/.

[3] S. Hochreiter, “The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 06, no. 02, pp. 107–116, 1998. [Online]. Available: https://doi.org/10.1142/S0218488598000094.

[4] Y. Huang *et al.*, “Artificial Intelligence in The Diagnosis of COVID-19 Challenges and Perspectives,” *International Journal of Biological Sciences*, pp. 1–7, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33907522/.

[5] Tanaka and Aranha, “Data Augmentation Using GANs,” *Proceedings of Machine Learning Research*, pp. 1–16, 2019. [Online]. Available: https://arxiv.org/pdf/1904.09135.pdf/.

[6] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *CoRR*, vol. abs/1106.1813, 2011. arXiv: 1106.1813. [Online]. Available: http://arxiv.org/abs/1106.1813.

[7] G. He Bai, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” *IEEE World Congress on Computational Intelligence*, pp. 1322–1328, 2008. [Online]. Available: https://www.researchgate.net/publication/224330873_ADASYN_Adaptive_Synthetic_Sampling_Approach_for_Imbalanced_Learning.

[8] Wang and Xiao, “Lychee Surface Defect Detection Based on Deep Convolutional Neural Networks with GAN-Based Data Augmentation,” *Agronomy*, vol. 11, 2021. [Online]. Available: https://www.mdpi.com/2073-4395/11/8/1500.