DEAKIN
UNIVERSITY AUSTRALIA

Transformer based COVID-19 detection using X-ray images

The Coronavirus Disease 2019 (COVID-19) pandemic keeps on spreading across the globe and the expeditious spread of COVID-19 since its outbreak has impelled many nations healthcare frameworks & economy to the edge of breakdown. Therefore, to suppress the spread of the disease and minimize the ongoing expenditure on the healthcare system, accurate identification & isolation of COVID-19 positive individuals for treatment is vital.

This research review discusses novel methods to curb the problem of COVID-19 cases using Transformers. We first discuss about the current methods employed in COVID-19 detection. Afterwards, we will discuss about the use of CNNs in COVID-19 detection and later we will discuss various transformer models, tools & their optimization techniques applied in detecting COVID-19 cases. Finally, we will provide conclusion.

Reverse Transcription Polymerase Chain Reaction (RT-PCR) has been the golden standard for diagnosing COVID-19 due to its accuracy & authenticity but is expensive and requires trained professionals, laboratory, and RT-PCR kit for COVID-19 detection & analysis which is time-consuming & arduous. Medical images such as chest X-rays (CXR) are crucial to confirming a COVID-19 diagnosis, as they provide accurate laboratory test as visual evidence to physicians and radiologists while being readily accessible in most healthcare systems. In the last few years, researchers have exhibited utilization of Deep Learning (DL) methods like Convolutional Neural Network (CNN) on chest X-ray (CXR) for COVID-19 detection which speeds up the COVID-19 diagnostic process, but CNN methods fail to capture the global context due to their inherent image-specific inductive bias. Therefore, an advanced Deep Learning model, Vision transformers using transformer architecture for images was proposed by Dosovitskiy et al. (2021) based on original self-attention-based architecture, in particular Transformers (Vaswani et al., 2017). The rise of Vision Transformers has likewise given a significant establishment to the development of vision models. ViT models presents a new state-of-the-art on Image Recognition with a model that, even fully depending on self-attention, is able to present performances on par with current state-of-the-art models.

The following review of research literature is based on detecting & classifying COVID-19 patients using Chest Radiography (X-ray) images, using different transformers conducted over the last two years, from 2020-2022 and provide a summary and statistical data analysis on the basis of performance metrics & results with previously implemented state-of-the-art methods using Transformers. This review of research literature excludes the use of transformers for Natural language processing (NLP) tasks as we are mainly dealing with applications utilizing transformers in computer vision domain for COVID-19 detection specifically.

Convolution Neural Network (CNN) are quite popular among the research community of AI in medicine as they produced best classification accuracy as compared to classification techniques like Artificial Neural Network (ANN), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) for COVID-19 detection using X-ray images. A work by Maram Mahmoud A. Monshi (2021) proposed CovidXrayNet, an optimized CNN model based on EfficientNet-B0 utilizing data augmentation & CNN hyperparameter tuning for detecting COVID-19 from CXRs in terms of validation accuracy. The author achieved state-of-the-art accuracy of 95.82% on the COVIDx dataset in the three-class classification task (COVID-19, normal or pneumonia) with only 30 epochs of training. They further classified that CovidXrayNet is still at a research stage and currently not suitable for direct clinical diagnosis.

Notwithstanding, existing Deep Convolutional Neural Network (CNN) methods neglect to capture the global context because of their inherent image-specific inductive bias. As of late, different investigations utilized the Vision Transformer (ViT) models for COVID-19 detection rather than CNN architectures. The challenge in medical images for ViT approaches comes in form of long-range dependencies and multi-modality. In ARNAB KUMAR MONDAL et al. (2021), they proposed a vision transformer based deep neural classifier, xViTCOS (Explainable Vision Transformer Based COVID-19 Screening Using Radiography). They employed a multi-stage transfer learning approach to address the problem of need for large-scale data & Gradient Attention Rollout algorithm (For Explanability). They demonstrated the viability of the proposed structure in distinguishing COVID-19 positive cases from non-COVID-19 Pneumonia and Normal control utilizing both chest CT scan and X-ray methodology, through several trials on benchmark datasets where they achieved an overall accuracy of 98.1% & recall (Sensitivity) of 96%. But xViTCOS-CT fails to predict the ground truth non-COVID-19 Pneumonia with confidence.

The image classification utilizing transformers can be optimized via transfer learning, data augmentation & hyperparameter tuning of the models. In Mohamed Chetoui & Moulay A. Akhloufi (2022), several ViT models (ViT-B16, ViTB32, and ViT-L32) were fine-tuned for the multiclass classification problem (COVID-19, Pneumonia and Normal cases). Data augmentation was applied during training, hyperparameter tuning of ViT models were done & visualized the signs detected by ViT by using the attention map of the best model (ViT-B32). Furthermore, demonstrated that the obtained results outperformed comparable state-of-the-art models for detecting COVID-19 on CXR images using CNN architectures. They report an accuracy of 96% & Recall (Sensitivity) of 96%. While they concluded that the recommended model would have an even better performance if compared to a manual interpretation by radiologists. Koushik Sivarama Krishnan & Karthik Sivarama Krishnan (2021) proposed a fine-tuned vision transformer (ViT-B/32) for detecting COVID-19 on chest X-rays. Noise is a key factor in radiography that affects the model's performance. For distinguishing COVID-19 cases from Viral Pneumonia, Lung Opacity, and Healthy chest X-rays, they acquired an accuracy score of 97.61%, Precision score of 95.34%, Recall (Sensitivity) score of 93.84% & F1-Score of 94.58% using ViT-B/32 transformer.

Debaditya Shome et al. (2021), used a constructed three-class data set of 30 K chest X-ray pictures for COVID-19 detection using Vision Transformer (ViT L-16) for healthcare. They designed a COVID-19 detection pipeline utilizing the Vision Transformer model and fine-tuned it on their dataset with a custom MLP block, resizing & interpolation of images, data augmentation techniques (random rotation, width shift, height shifts, and flipping). For better model interpretability and simplicity of diagnosis, they created Grad-CAM-based visualizations of COVID-19 movement in the lungs, which helps the diagnosis process for medical services. They report high performance with accuracy and AUC score as high as 98% and 99%, respectively for classifying COVID-19 from normal chest X-rays in the binary classification task. Furthermore, distinguishes COVID-19, normal, and pneumonia patient's X-rays with an accuracy of 92% and AUC score of 98% in the multi-class classification task. The only disadvantage with their model was that the model results in overfitting beyond 2 dense layers. The conducted study might be very important in regions where quick testing is inaccessible, and it might likewise be utilized as a second screening method after the standard RT-PCR test to check that any true negative or false positive cases don't happen.

The self-attention transformer-based approach is of fundamental importance for the methods intent to analyze the COVID-19 in CT scan images. In Fozia Mehboob et al. (2022) proposed a novel architecture of vision transformers using self-attention Transformer based diagnosis approach for the diagnosis of COVID-19 using 3D CT Slices. The authors employed image segmentation using transformer encoder/ decoder architecture, layer normalization & data augmentation (image flipping (horizontal), resizing (image size) and rotation) for optimizing image classification. They compared the performances of self-attention transformer-based approach with CNN and Ensemble classifiers for diagnosis of COVID-19 using Brazilian dataset (SARS-COV2 CT scan dataset) & COVID-19 (HUST-19) CT scan dataset. Upon evaluation, the authors achieved an accuracy of 98% on Brazilian dataset & 99.7% on multi-class Hust19 CT scan dataset. The proposed transformer vision approach can predict the quantification of COVID-19 based on the pixel values in the long-range relation-based maps which can provide valuable assistance to specialists in decision making with respect to the assessment of the severity of the COVID-19.

Swin Transformer (Ze Liu et al., 2021) outstandingly refreshed past records on object detection and semantic segmentation benchmarks, providing the community prominent certainty that the Transformer structure will turn into the new standard for visual modeling. In Juntao Jiang & Shuyi Lin (2021), the authors proposed a method which combined Swin Transformer (Swin-B) and Transformer in Transformer (small) for classifying COVID-19, Pneumonia and Normal (healthy) cases using chest X-ray images and did model ensemble by using weighted average method. The authors utilized data augmentation (horizontal flipping, and the rotation or translation), applied label smoothing in loss & performed hyperparameter tuning of the model for optimizing the model. The authors report achieving an accuracy of 94.75% for detecting COVID-19 cases.

ViT does not generalize well when trained on insufficient amounts of data while Data-Efficient Image Transformer (DeiT) almost follows the same architecture as ViT but follows a teacher-student strategy with distillation token & provides better performance as compared to competitive convnet models and most state-of-the-art ViT models. In Mohamad Mahmoud Al Rahhal et al. (2022), presented a novel architecture based on a Data-Efficient Image Transformer (DeiT) architecture which is an improved version of Vision Transformer (ViT) for COVID-19 detection using CT (SARS-CoV-2 CT) & X-ray images (COVIDx). The authors employed a Siamese encoder that implements a distillation technique to classify original and augmented images. Heat maps were utilized which showed the progression of focus areas over network layers, similar to X-ray or CT images. Upon evaluation, the authors achieved an accuracy of 94.62% on CXR dataset & 99.13% accuracy on CT dataset which showed the model's robustness under limited training data.

Pyramid Vision Transformer (PVT) proposed by Wenhai Wang et al. (2021) overcomes the challenges of porting Transformer to different dense prediction tasks. PVT can be trained on dense partitions of the image to accomplish high output resolution and can reduce computations of large feature maps using a progressive shrinking pyramid structure. In Xiaoben Jiang et al. (2022) proposed a new variant of pyramid vision Transformer (MXT) for multi-label chest X-ray image classification which can capture both short and long-range visual information through self-attention. The authors employed multi-layer overlap patch (MLOP) embedding, class token Transformer block and multi-label attention (MLA) for more effective processing of multi-label classification. The authors evaluated MXT on a large-scale CXR dataset (Chest X-ray14) with 14 disease pathologies & Catheter dataset which resulted the highest mean AUC score of 83.0% on the Chest X-ray14 dataset and 94.6% on the Catheter dataset. While the heatmaps generated from the Model-D were discrete and blurry (i.e., CXR image with Atelectasis, Pneumothorax), and location of the lesion regions is in error which presented the limitations of the presented model. Overall, MXT can help radiologists in diagnoses of lung diseases and check the position of catheters, which can lessen the work strain of medical staff.

After reviewing the relevant literature, Transformer's extraordinary success in NLP has been explored in the computer vision domain & presents a crucial step in the research direction. Model architectures in the related work of adopting transformer in computer vision very well may be utilized in a pure Transformer manner or in a hybrid manner by combining with CNNs. Utilizing convolutions with transformers can enhance the performance & efficiency of the models. In spite of the viability of deep learning-based frameworks in COVID-19 identification, models that are trained on small datasets will lead to improper generalization because of which the model could perform inadequately in real world scenario and having a small test set could bring about missing out on false positives or negatives. Transformers are still profoundly data-dependent and is undoubtedly a significant step towards research in computer vision domain through which we can save a lot of time, cost & labour in medical industry for COVID-19 detection.