Vision Transformer for Covid-19 X-Ray Image Classification

CS 598 Deep Learning For Healthcare: Project Report

Kyle Maxwell (Computer Science, UIUC, kylem6@illinois.edu)
Tom Phelan (Computer Science, UIUC, tphelan3@illinois.edu)
Ben Roberson (Computer Science, UIUC, br13@illinois.edu)
Bingying Yong (Computer Science, UIUC, byong2@illinois.edu)

ABSTRACT

A new strain of coronavirus named Severe Acute Respiratory Syndrome Coronavirus 2 (Covid-19) was identified in Wuhan, China in late 2019. Covid-19 was officially named as a pandemic on March 11, 2020 [1]. A formal diagnosis usually requires using a PCR test [2], however the results can be unreliable [3]. The radiographic characteristics of Covid-19 pneumonia are atypical but require careful observation and diagnosis from a radiologist [4]. So, with the limited amount of medical resources, it would be advantageous to develop a deep learning model that is able to effectively diagnose Covid-19 using x-ray images.

In this study, we achieved this by using transfer learning with a pre-trained Vision Transformer (ViT) model, then fine-tuned using a dataset of 1685 x-rays (1489 Covid-19 negative and 196 Covid-19 positive). Our model achieved an accuracy of 0.917, precision of 0.642, and recall of 0.915.

Our ViT model achieved comparable or better performance across all metrics compared to the ResNet-18 model. However it also saw an increase in training time. Potential improvements to our model could be to utilize additional state of the art innovations to the vision transformer architecture, such as DeepViT [12] and CrossViT [13].

Keywords

Covid-19, X-ray, Image Classification, Medical Diagnosis, Pandemic, Vision Transformer, ViT

1. INTRODUCTION

Coronaviruses are a large family of viruses that are common and typically only cause mild illnesses in humans. In late 2019, a new strain of the disease was identified in Wuhan, China. This novel coronavirus has since been named Severe Acute Respiratory Syndrome Coronavirus 2 (Covid-19). On March 11, 2020, the World Health Organization (WHO) officially assessed Covid-19 as a pandemic [1].

Common symptoms of Covid-19 include fever, cough, fatigue, dyspnea, and myalgia, and a formal diagnosis is usually confirmed using Reverse Transcription Polymerase Chain Reaction (RT-PCR) on nasopharyngeal and throat swabs [2]. The performance of PCR testing (SARS-CoV-2 rtRT-PCR) can be unreliable. A study published in January 2021 found a false negative rate of 9.3% in its test cases [3]. Additionally, a great number of cases went undiagnosed before the disease was understood and proper testing was created.

A study done by Rousan et. al in 2020 found that almost half of the patients with Covid-19 had abnormal chest x-ray findings, with the most common abnormality being peripheral ground glass opacities (GGO) affecting the lower lobes. A case study in Korea also found that in PCR confirmed cases, 70% of the observed opacities were in consolidation [4]. These characteristics class Covid-19 pneumonia as an atypical pneumonia, and these changes can also be observed in other atypical pneumonias such as SARS and MERS. However, these radiographic changes can be subtle and require careful observation and diagnosis from a radiologist [4]. This creates constraint on the already limited medical resources available and could bottleneck the treatment process. So it would be advantageous to develop a diagnostic model based on existing x-ray technology, from both a treatment and an epidemiological perspective. Its application might be useful even in future outbreaks of a novel virus, If a prediction model can provide diagnosis with good accuracy using only an x-ray image, then this will allow healthcare professionals to stay ahead of the emerging disease and offer their patients proper care.

Since the start of the pandemic, there have been several promising studies done in the area of using x-ray images to diagnose Covid-19 cases. Earlier attempts mostly employed the classic neural networks, such as the pre-trained ResNet50 model used by Narin et al [5], and the Resnet50 plus SVM model used by Sethy and Behera [6]. Both studies achieved very accurate binary classification results (Covid-19 positive versus Covide-19 negative), with 96.1%

[5] and 95.4% [6] accuracy respectively. More recent attempts, however, have tried to innovate on the methodology used. Ozturk et al proposed a deep learning model that they called DarkCovidNet [7]. This is based on the architecture of an existing model called DarkNet-19, which uses 19 convolutional layers and 5 pooling layers in a CNN. With DarkCovidNet, the authors used fewer layers and filters, and also used LeakyReLU operations to capture fine-level details without zero-ing out the neurons. The result is a binary classification model that achieved an accuracy of 98% [7]. Khan et al also proposed a new neural network model called CoroNet, which is based on Xception architecture and pre-trained on ImageNet dataset, then trained end-to-end with Covid-19 x-ray data. The binary classification results of CoroNet reached an accuracy of 99% [8].

Outside of Covid-19 classification, researchers at Google have recently attempted to bring the success of Transformers to the image recognition domain [9]. Transformers have proven to be highly effective in the domain of NLP and machine translation. Since Transformers require linking each token in an input set to all the other tokens present (or each pixel in an image to all other pixels), they are exponentially computationally expensive relative to the size of the input. Therefore they were thought to be too impractical for image datasets. The authors introduced novel techniques such as dividing the image into a sequence 16x16 patches and using the full Transformer technology only within each patch, thus greatly reducing the computational complexity. The resulting model, which they called Vision Transformer (ViT), performed quite well in object recognition. The authors have also provided pre-trained models using common image datasets [9]. Since the ViT architecture is quite new, its use has not been fully explored. Therefore the purpose of our work is to apply the ViT model to the domain of Covid-19 detection using x-ray images.

2. METHODOLOGY

2.1 Dataset

The first data source that we will be using comes from a publicly available dataset compiled by Cohen et al [10]. This dataset has been manually aggregated from publications and various web based repositories and contains chest x-ray and CT images of patients who have either been positively diagnosed or were suspected of having Covid-19 or other viral and bacterial pneumonia (MERS, SARS, ARDS). Currently there are 949 images available from 472 patients. Out of these records, 584 were determined to be positive cases of Covid-19. Additional clinical attributes such as sex, age, survival, intubation status, etc. are also provided, although certain records may be missing some of this information [10].

Since the first data source contains predominantly x-ray images from the positive Covid-19 class, we supplemented the negative class using a second publicly available dataset from the kaggle repository "Chest X-Ray Images (Pneumonia)" [11]. A total of 1341 x-ray images were added, none of which demonstrate any evidence of Covid-19 or pneumonia.

Since our image data comes from a number of different international sources, there are some variations to the files. Therefore we performed the following pre-processing steps:

- Labelled the x-ray images as either Covid-19 positive (label = 1) or Covid-19 negative (label = 0)
- Only used x-rays showing a frontal posteroanterior view (PA view)
- Resize the images to 255x255
- Crop the images at the center and create an output image of size 244x244
- Normalize the three image channels in the range of [-1, 1], using mean of (0.5, 0.5, 0.5) and standard deviation of (0.5, 0.5, 0.5), which aligns with the pre-trained model's pre-processing.

Our final dataset contains a total of 1685 x-rays, with 1489 Covid-19 negative images (including normal and Covid-19 negative pneumonia) and 196 Covid-19 positive images.

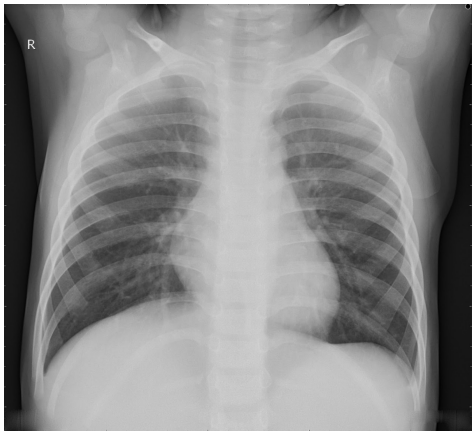

Figure 1: This image shows a Covid-19 negative x-ray image post data pre-processing.

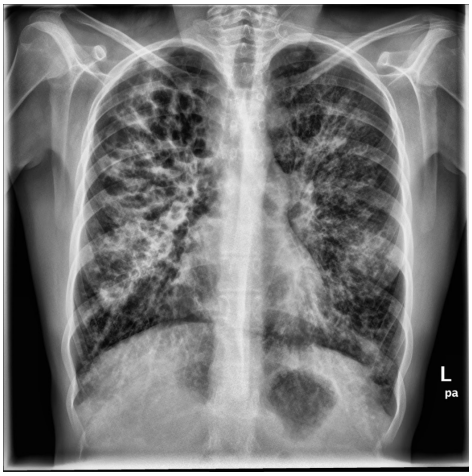

Figure 2: This image shows a Covid-19 positive x-ray image post data pre-processing.

2.2 Model

Visual Transformers (ViT) was developed by Dosovitskiu et al in 2020 to apply the transformer architecture to image classification. Traditionally in computer vision, attention was only used in conjunction with or used to replace certain components of convolutional networks (CNN). However ViT models do not necessarily rely on CNNs and instead directly apply transformers to sequences of image patches [9]. The ViT model splits the images into fixed size patches, then linearly embed each patch with position embeddings. This sequence is then fed into a standard transformer encoder. Finally, an extra learnable 'classification token' is added to the sequence in order to generate classification outcomes [9].

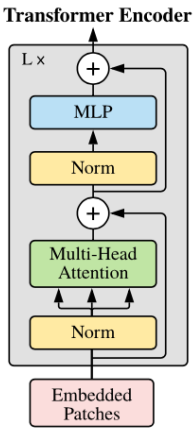

Figure 3: Multi-Head Attention Transformer Encoder Block [9]

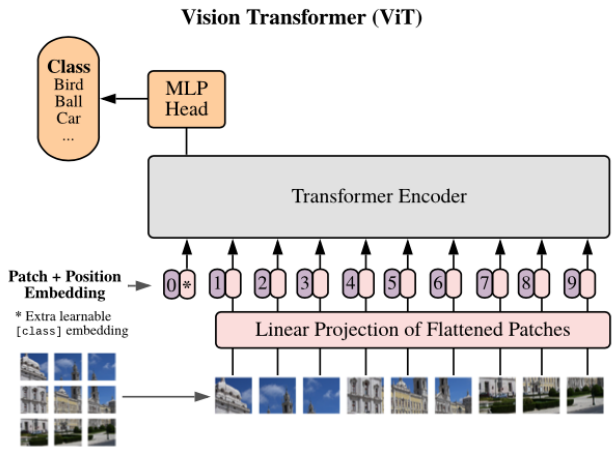

Figure 4: Vision Transformer Architecture [9]

The goal of our model is to classify an x-ray image as either Covid-19 positive or negative. We fine-tuned the model using the Covid-19 dataset described in the previous section using an 80-20 training validation split. Our model was built from a pre-trained ViT model (B_16) that has been trained on ImageNet-21K images.

We replaced the final linear multi-class classification layer end of the model to predict the covid-19 binary outcome. In order to accelerate training, we first froze all of the twelve multi-head attention transformer encoder blocks within the model. During each epoch, we unfroze one additional transformer encoder block such that in the preliminary epochs, only the last parameter blocks were trainable, and in the last epochs the entire ViT was able to be fine-tuned.

In training this model we used the CrossEntropyLoss function of pytorch and the SGD optimizer. Because the training dataset is imbalanced (about 7:1 negative to positive images), we weighted the loss function accordingly.

For comparison purposes, we also trained a standard ResNet18 model from the torchvision models subpackage, using the same dataset, the same pre-processing, and (where applicable) the same model adjustments.

Please see
https://github.com/BenDRoberson/COVID19-ViT-Project
for all code and data to replicate these results.

3. RESULTS

The input data was randomly split into two subsets, with 80% used for training and 20% used for testing. The models were trained for 12 epochs to convergence. The binary classification performance of our model has been captured using accuracy, precision, recall, ROC AUC, and F1 score.

	Our ViT Model	ResNet-18
Accuracy	.917	0.920
Precision	.630	0.560
Recall	.979	0.848
ROC AUC	.943	0.888
F1	.767	0.675

Table 1: Comparison of our ViT model results against the ResNet18 model

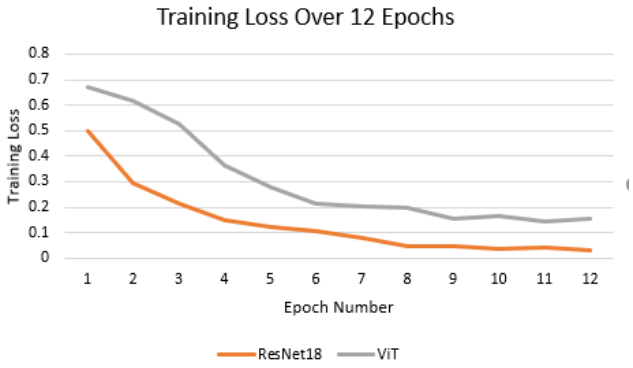

Figure 5: Training loss comparison over between our ViT model results and the ResNet18 model results

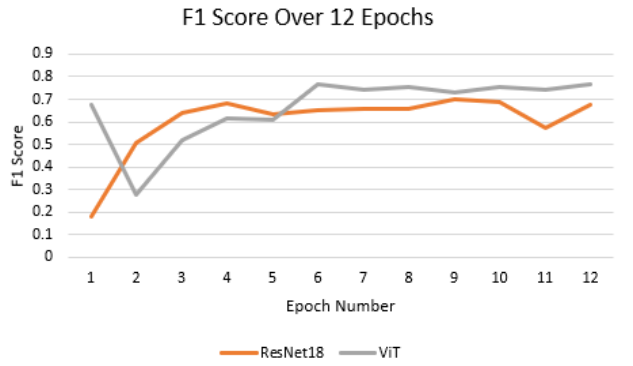

Figure 6: F1 score comparison over between our ViT model results and the ResNet18 model results

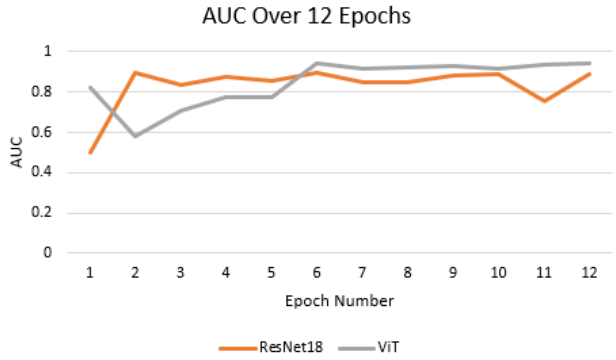

Figure 7: AUC comparison over between our ViT model results and the ResNet18 model results

4. DISCUSSION

Using the ViT model, we were able to achieve comparable accuracy but better performance on all the other metrics when compared to the previous state-of-the-art ResNet18 model. One disadvantage to the ViT model is the increased training time - in our testing it took approximately three times as long to train an equivalent number of epochs of ViT as it did of ResNet18. This is to be expected because of the more complicated architecture, which has a far greater number of trainable parameters. However, there are further opportunities to use the ViT model, such as the ability of ViT to take higher resolution images by increasing the number of patches and remapping them [9].

Additionally, in recent months there have been a multitude of transformer improvements that could be combined with ViT such as DeepViT [12], a method for increasing the depth of ViT, or CrossViT [13], a method that trains two vision transformers that process the image at different scales, cross attending to one every so often. However, both of these methods are quite computationally intensive, and would require a larger dataset to be effective, so we have chosen to not explore them in this paper. Lastly, we could see a lift in performance by pre-training on a much larger x-ray image dataset built for non-covid multi-class classification models.

An extension of this project might be to use the saliency map for model interpretability. Since the model already uses the attention mechanism, we could use that mechanism to explain why an image was categorized as the positive or negative class. In other words, we might be able to make a predictive diagnosis based on specific identifiable anomalies of an x-ray image, similar to how a healthcare professional would do it. The authors of the ViT paper proposed a method of doing this by using the first layer of the model, which projects the 16x16 pixel image patches into a low-dimensional space. This projection has an embedding layer added on to it that encodes the

similarity of the patches; this can be used to identify the areas of greatest attention on the original image

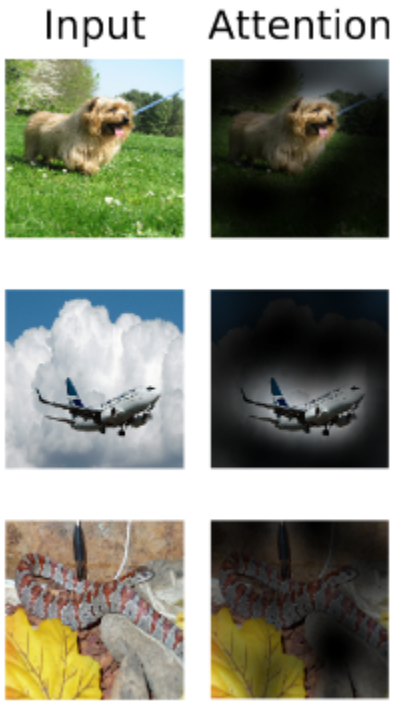

Figure 8: Example images showing attention from the output token to the input space, which highlights the regions that are semantically relevant for classification [9]

Another extension might be to take the predictions a step further and attempt to predict the actual mortality or disease severity of the patient.

5. CONCLUSION

ViT is a promising new technique for X-Ray image classification. We were able to achieve superior results as compared to a Resnet18 model, and with an accuracy of over 90% this is a technology with potential applications in the clinical environment. Future work in this area can focus on incorporating recent advances in this emerging technology, and using the saliency map for highlighting important areas of the image.

6. REFERENCES

[1] Canada, P. H. (2021, March 05). Government of Canada. Retrieved March 28, 2021, from https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html

[2] Rousan, L.A., Elobeid, E., Karrar, M. et al. Chest x-ray findings and temporal lung changes in patients with COVID-19 pneumonia. BMC Pulm Med 20, 245 (2020). https://doi.org/10.1186/s12890-020-01286-5

[3] Kanji, J.N., Zelyas, N., MacDonald, C. et al. False negative rate of COVID-19 PCR testing: a discordant testing analysis. Virol J 18, 13 (2021). https://doi.org/10.1186/s12985-021-01489-0

[4] Cleverley J, Piper J, Jones M M. The role of chest radiography in confirming covid-19 pneumonia BMJ 2020; 370 :m2426 doi:10.1136/bmj.m2426

[5] Narin, A., Kaya, C., & Pamuk, Z. (2020). Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. arXiv preprint arXiv:2003.10849

[6] Sethy, P.K.; Behera, S.K. Detection of Coronavirus Disease (COVID-19) Based on Deep Features. Preprints 2020, 2020030300 (doi: 10.20944/preprints202003.0300.v1)

[7] Ozturk, T., et. al. Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images. Computers in Biology and Medicine 121. (April 2020). https://www.researchgate.net/publication/340935440_Automated_Detection_of_COVID-19_Cases_Using_Deep_Neural_Networks_with_X-ray_Images

[8] Khan, A., et. al. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Computer Methods and Programs in Biomedicine 196. (November 2020). https://www.sciencedirect.com/science/article/pii/S0169260720314140

[9] Dosovitskiu, A., et. al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929

[10] Cohen, J.P., Morrison, P. & Dao, L. (2020) COVID-19 Image Data Collection. arXiv:2003.11597, https://github.com/ieee8023/covid-chestxray-dataset, 2020

[11] Chest X-ray images (pneumonia). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. Last Accessed: April 11, 2021

[12] Zhou, D, et. al. (2021) DeepViT: Towards Deeper Vision Transformer. https://arxiv.org/abs/2103.11886

[13] Chen, C, et. al. (2021) CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. https://arxiv.org/abs/2103.14899

---

7. About the authors:

Kyle Maxwell is a Machine Learning and Software Development Engineer at GoDaddy in San Francisco, CA. He can be found on linkedin.com/in/kylemaxwell.

Tom Phelan is a Research Data Analytics Engineer at MHealth Fairview Hospitals in Minneapolis Minnesota. His public projects can be viewed at https://github.com/tphe.

Ben Roberson is a Senior Modeling Analyst at GEICO in Washington DC. He can be found on LinkedIn here: https://www.linkedin.com/in/benroberson123/.

Bingying Yong is a business consultant with Varicent Software in Toronto, Canada. She can be found on LinkedIn at https://www.linkedin.com/in/byyong.