Classification of Virus-Infected Chest X-Ray Dataset

Jinyu Hou, Ziyi Zhou, Zhiyuan Lu
Department of Computer Science
University of Toronto

Abstract

Virus-infected lungs have altered structure compared to normal lungs that could be observed through chest X-ray imaging due to pulmonary fibrosis. Our project aims to simulate the diagnostic process by classifying chest X-ray images into virus-infected lung images and normal lung images. We fine-tuned two ImageNet-21k pre-trained model: BiT-M-R50x1 (Kolesnikov et al., 2020) and R50+ViT-B/16 (Dosovitskiy et al., 2020) on the Chest X-Ray dataset.

1 Introduction

COVID-19 has become one of the biggest threat to the community since 2020. As a disease caused by virus infection of the lung, chest X-ray images can be an efficient tool for its diagnosis. Our project will apply and benchmark two state-of-art deep learning models on this classification task of separating virus-infected lung images from normal lung images.

The main task of this project is to perform transfer learning with BiT-M-R50x1 (Kolesnikov et al., 2020) and R50+ViT-B/16 (Dosovitskiy et al., 2020) on the COVID-ChestXRay dataset (Cohen et al., 2020a). Transfer learning allows us to use complicated models pretrained on general datasets such as Imagenet-21k for tasks of different domains. We plan to first fine-tune the two image classification models and compare their performance in terms of accuracy. Then, we would analyse and compare their decision policy.

2 Related works

Chest X-ray images have been the focus of many researches that apply machine learning models for the automatic diagnosis. Since the outbreak of COVID-19, different machine learning models have been developed.

Diagnostic models with classic image recognition models Most of the papers considering COVID-19 and pneumonia diagnosis used classic models for image recognition, including ResNet-18 (Tartaglione et al., 2020), ResNet-50 (Ghoshal and Tucker, 2020), DenseNet-121 (Zhang et al., 2021), and EfficientNet (Luz et al., 2021). ResNet and DenseNet architectures showed better results than the others, with accuracies ranging from 0.88 to 0.99 (Roberts et al., 2021).

Transformers Transformers (Vaswani et al., 2017) have become the model of choice in many natural language processing (NLP) tasks. Models are often pre-trained on a large text corpus and then fine-tuned for a specific task. However, there have been experiments that directly apply a Transformer to images, where an image is splitted into patches and a sequence of linear embeddings of these patches are the input to a Transformer (Dosovitskiy et al., 2020). There have been applications of transformers to computer vision tasks, but few used medical imaging data.

Preprint. Under review.

3 Methods

Big Transfer The first baseline model, "Big-Transfer(BiT)", was released in 2020 by Google Research in Zürich, Switzerland. The model we use (BiT-M-R50x1) is based on the ResNet-50x1 architecture and was pretrained on the ImageNet-21k dataset (Kolesnikov et al., 2020).

Vision Transformer The second model, "Vision Transformer" (ViT), was also released in 2020 by Google Research. The model we use (R50+ViT-B/16) is a transformer with ViT-B/16 on top of a ResNet-50 (BiT) backbone. Simulating how transformers work in NLP tasks, the ViT model divides the intermediate feature maps from BiT into patches and transform them into a sequence of linear embeddings. These patches are equivalent to BERT's tokenized words in NLP tasks. This enabled us to apply self-attention to images without increasing the computational cost for too much, compared to pixel-to-pixel attention (Dosovitskiy et al., 2020).

Although Vision Transformers lack the inducive bias inherent in CNN models and do not generalize well in small datasets, after pre-trained at sufficient scale (ImageNet-21k dataset), Vision Transformers can outperform the state-of-art CNN model.

Dataset The dataset we use was obtained from Kaggle. The original source of COVID-ChestXRay was collected from the Guangzhou Women and Children Medical Center, Guangzhou, China in 2020. There are a total of 5,857 chest X-ray images in the dataset (Cohen et al., 2020b). The images are divided into the following class hierarchy:

Table 1: Class Hierarchy of COVID-ChestXRay Dataset

Label_0	Label_1	Label_2	Image Count
Normal			1576
Pneumonia	Stress-smoking	ARDS	2
Pneumonia	Virus		1493
Pneumonia	Virus	COVID-19	58
Pneumonia	Virus	SARS	4
Pneumonia	Bacteria		2772
Pneumonia	Bacteria	Streptococcus	5

Since it is obviously an imbalanced data that some of the sub-classes have only a few cases, we decided to classify only normal vs. all virus infected cases. Therefore, the resulting dataset has two classes: 0 - Normal (1576 cases), 1 - Virus-infected (1555 cases). The images are resized and center-cropped during preprocessing.

Label Smoothing Cross Entropy Loss Label smoothing is a way to calibrate our model so that its predicted probabilities could be close to the test accuracy. This could prevent the model from begin overfitting by introducing noise into the target label when training. The equation is as follows:

$$t_{ls} = (1 - \alpha) * t + \alpha / K$$

in which t_{ls} represents the smoothed label, t represents its original one-hot encoded label, α is the smoothing hyperparameter and K is the number of label classes. (Müller et al., 2020) Then t_{ls} is used to calculate the cross-entropy loss for gradient descent. This is applied in the ViT model.

4 Experiment

4.1 Accuracy Comparison

We compare the performance of ViT-B/16 (ViT), R50+ViT-B/16 (hybrid) and BiT-M-R50x1 (BiT) model on the COVID-ChestXRay Dataset after fine-tuning. All the models are pre-trained on the ImageNet-21k dataset.

According to the test data in Dosovitskiy et al. (2020) and Kolesnikov et al. (2020), the ViT model outperforms BiT model on transfer performance on various datasets. The R50+ViT hybrid model slightly outperforms ViT, but the difference vanishes in larger models. We used small sized models in our experiment, so the hybrid model was expected to perform slightly better than ViT on the COVID-ChestXRay Dataset.

In our experiment, all the models are fine-tuned and tested on Google Colab single GPU environment, and we use the same simple image preprocessing method on the COVID-ChestXRay Dataset for each model. The BiT and ViT/Hybrid models are trained with different optimizer, loss function and learning rate schedules (Table 2).

Table 2: Hyperparameter Choices

	BiT	ViT/Hybrid
optimizer	Adam (lr=0.1)	SGD (momentum=0.9, lr=0.01, weight decay=0.0001)
loss function	CrossEntropyLoss	LabelSmoothingCrossEntropy (smoothing=0.1)
learning rate schedule	1/t decay	setup decay

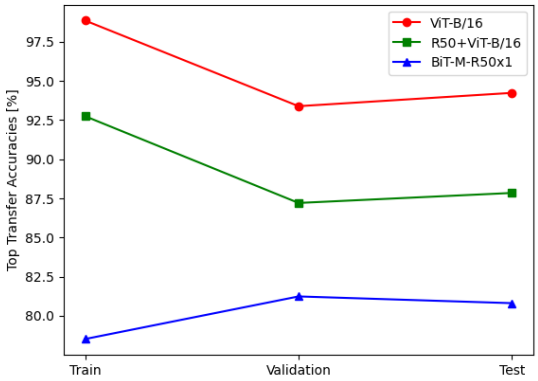

Figure 1: Transfer to COVID-ChestXRay Dataset. ViT / R50+ViT model outperforms BiT model on the COVID-ChestXRay Dataset after fine-tuning. The R50+ViT model achieved slightly lower accuracy than ViT.

Figure 2 shows the results. After pretrained on the ImageNet-21k dataset, the ViT / R50+ViT model outperforms BiT model on the COVID-ChestXRay Dataset after fine-tuning. The result is the same when performed on other various downstream datasets, according to Dosovitskiy et al. (2020). This demonstrates that although Vision Transformers lack the inducive bias inherent in CNN models and do not generalize well in small datasets, after pre-trained at sufficient scale (ImageNet-21k dataset), Vision Transformers can outperform the state-of-art CNN model (Dosovitskiy et al., 2020).

The R50+ViT hybrid model achieved slightly lower accuracy than ViT, which is not what we expected. The patch embedding projection in R50+ViT hybrid model is applied to patches extracted from the BiT feature map, and the convolutional local feature processing was expected to assist ViT performance.

The training curve for BiT, ViT and R50+ViT is included in the **6.5** section of the appendix.

4.2 Activation Map

Gradient-weighted Class Activation Mapping (Grad-CAM) produces activation mapping which highlights regions in the input image which is critical for the network to make decision. The way it works is to visualize the gradient of the upper convolutional layer that produces the corresponding target label (Selvaraju et al., 2019). It is useful to investigate the way a neural network architecture works. The heatmaps produced are as follows:

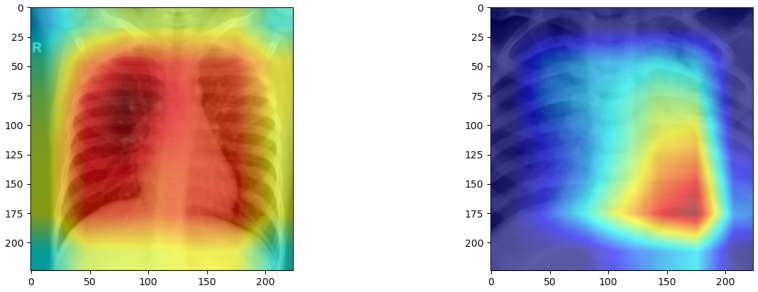

Figure 2: Activation Map of BiT-M-R50x1 (left - normal, right - virus-infected)

We only included the activation heatmap for the BiT(ResNet50) model because this model uses convolutional layer for its prediction. The ViT model only uses convolutional layer for lower-level feature extraction and applies self-attention in its upper layers for its prediction, which makes it trivial to analyse any convolutional layers. In the result, the heatmap for "normal" class activation focuses on the throughout pattern of the lung while the heatmap for "virus-infected" class activation focuses more on a specific area in the X-ray image. This matches the current medical knowledge that virus-infection produces opaque fibrous areas in the lung which could be observed through chest X-ray and that normal lung would have a more transparent and clear throughout pattern in the image in comparison.

4.3 Activation Maximization

Activation maximization is a technique in which the parameters in the trained model is fixed and we runs gradient ascent on the activation of the final convolutional layer with respect to the input image for a specific output class (Ogura and Jain, 2020). This could help us know how the model works by producing the pattern which activates the model prediction for a specific class for the most. In other words, it produces the most "typical" pattern of the specific class. The result is as following: (0 - Normal, 1 - Virus-infected)

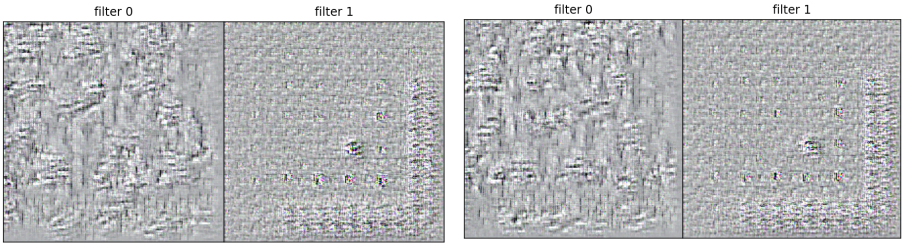

Figure 3: Activation Maximization of BiT-M-R50x1(left) and R50+ViT-B/16(right)

From the result, we could see that the two models produced similar results, which means that the two models have somehow similar decision policies. In the resulting images with filter 0 (normal), clear contraction of light and dark areas exists. This pattern resembles the partial looking of the ribs. The image produced with filter 1 (virus-infected) has a more fine-grained pattern with a smaller contraction which resembles partial looking of pulmonary fibrosis.

5 Conclusion

In summary, we have performed classification on chest X-Ray images by using the ResNet and Vision Transformers model on the COVID-ChestXRay Dataset after fine-tuning. Our results demonstrate that after pre-trained at sufficient scale (ImageNet-21k dataset), Vision Transformers (ViT-B/16, R50+ViT-B/16) can outperform the state-of-art CNN model (BiT-M-R50x1).

Given the accuracy from the models, they can potentially aid in the diagnosis of virus-infection caused pneumonia. However, as any clinically applied algorithms have to be dealt with caution, we need further experiments to evaluate its utilization in clinical practice.

References

Cohen, J. P., Morrison, P., and Dao, L. (2020a). Covid-19 image data collection. *arXiv 2003.11597*.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. (2020b). Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Ghoshal, B. and Tucker, A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big transfer (bit): General visual representation learning.

Luz, E., Silva, P., Silva, R., Silva, L., Guimarães, J., Miozzo, G., Moreira, G., and Menotti, D. (2021). Towards an effective and efficient deep learning model for covid-19 patterns detection in x-ray images. *Research on Biomedical Engineering*.

Müller, R., Kornblith, S., and Hinton, G. (2020). When does label smoothing help?

Ogura, M. and Jain, R. (2020). Misaogura/flashtorch: 0.1.2.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., and Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M., and Grangetto, M. (2020). Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data. *International Journal of Environmental Research and Public Health*, 17(18):6933.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhang, R., Tie, X., Qi, Z., Bevins, N. B., Zhang, C., Griner, D., Song, T. K., Nadig, J. D., Schiebler, M. L., Garrett, J. W., Li, K., Reeder, S. B., and Chen, G.-H. (2021). Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence. *Radiology*, 298(2):E88–E97.

6 Appendix

6.1 Contribution

Jinyu Hou: Write report, BiT fine-tuning and analysis
Ziyi Zhou: Write report, ViT fine-tuning and analysis
Zhiyuan Lu: Write report, test code

6.2 GitHub Repo Link

`https://github.com/jinyu-hou/csc413-project`

6.3 Software Setup

All the models are trained on Google Colab with CUDA. All the implementations are based on Python 3.7 and PyTorch 1.0.2. We applied the implementation of Big Transfer and Vision Transformer by Google Research.

6.4 Cited Project Link

https://github.com/google-research/big_transfer
https://github.com/google-research/vision_transformer
https://github.com/ieee8023/covid-chestxray-dataset
https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset
https://github.com/rwightman/pytorch-image-models
https://github.com/tanjimin/grad-cam-pytorch-light
https://github.com/MisaOgura/flashtorch

6.5 Training Curves

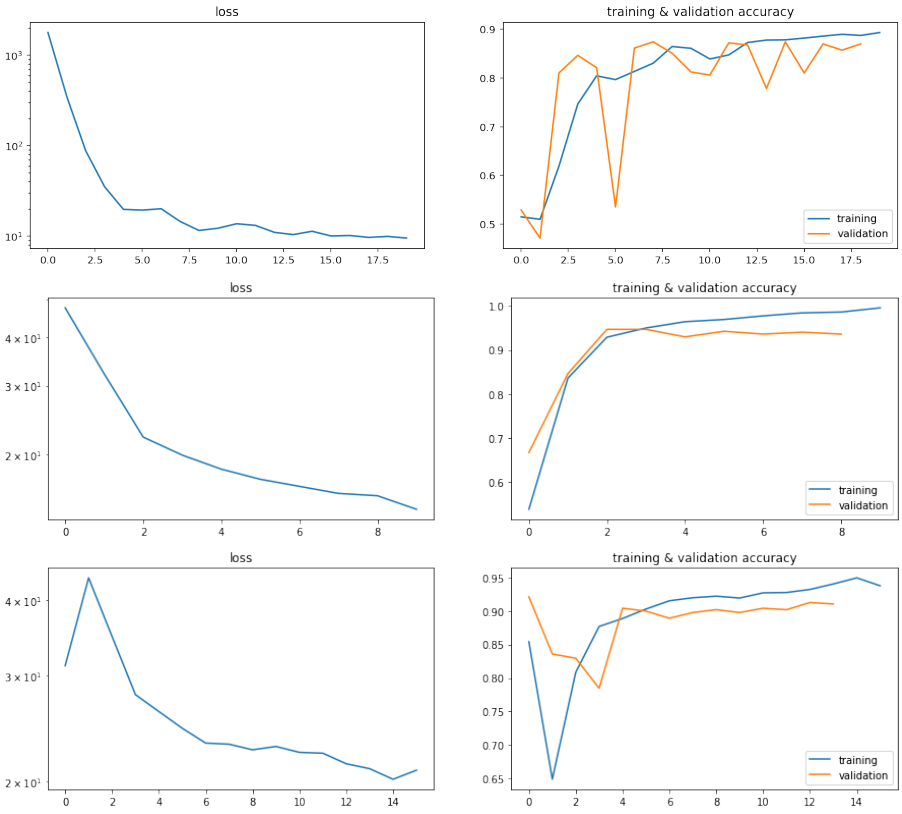

Figure 4: Loss, training and validation accuracy of BiT, ViT and R50+ViT