The Advantages and Disadvantages of Applying Transformers in Medical Images

Chun-Fu Yeh, Yue Liu, Nagul Ulaganathan
University of Pennsylvania
cfyeh@seas.upenn.edu, liuyue1@seas.upenn.edu, nagul@seas.upenn.edu

Abstract

The deep learning techniques, particularly convolution neural networks (CNNs), have achieved great success in automating workflows with medical images. Despite high performances with CNNs, the interpretability of them for medical images is of concern, and an additional module, such as GradCAM, needs to be manually inserted into CNNs. With the advent of ViT and Swin Transformer, which are the transformer-based networks for computer vision tasks, the concern in the interpretability of CNNs may be mitigated. Our results show that both ViT and Swin Transformer could achieve similar performances with the CNNs in terms of the transfer learning in the medical image classification task. Furthermore, the built-in attention module provides better interpretability for medical images. Compared to GradCAM, the attention map has a more spread out pattern, several local regions jointly contributing to the prediction. This would be the advantages of the transformer-based models over CNNs. Further studies should be conducted to address the vulnerability of the transformer-based models in hyperparameter tuning in medical image tasks, particularly for high resolution medical images.

1. Introduction

With the thriving and rapid-growing development of deep learning models, the medicine and healthcare industry has been a targeted and popular domain for the application of deep learning. In the recent three years, there have been many state-of-the-art convolutional neural networks coming out, competing with each other, based on the performance in the ImageNet dataset. However, as the CNNs become more complex, the researchers started to consider another framework other than CNNs to handle various kinds of image tasks. The most famous one is the transformer-based model, which is the dominant framework for natural language processing (NLP) [10]. Different from CNNs which detect spatially invariant and local features, the transformer-based model aims to learn the global context of a given input with the built-in multi-head attention module. This model framework may be helpful to the applications in medicine and healthcare as the medical professionals would take both local features and global contexts into consideration when making diagnosis. Based on the recent studies, the transformer-based model has achieved the performance on par with the CNNs on the ImageNet. However, the applications of transformer-based models in medical images are still in the early stage. In this study, we would like to investigate what are the advantages and disadvantages of applying transformer-based models on the medicine and healthcare industry based on the performance in medical image classification with chest x-rays.

Contributions.

- We investigate the advantages and the disadvantages of the transformer-based models in terms of the transfer learning in the classification task with medical images.
- The baseline result shows that the transformer-based models could achieve similar performances with those in the convolution neural networks if the size of the medical image is the same with the input formats in their pretrained models.
- The attention map in the transformer-based models is the build-in and learned weights, containing more information about the model interpretability (usually in a spread-out pattern and several local regions jointly contribute to the prediction).
- Our result with higher resolution image (512 x 512) shows that the convolution neural networks are still the first to be considered for transfer learning as it takes time for the transformer-based networks to re-wire the global interactions between the patches and they more sensitive in hyperparameter tuning.

2. Related works

The transformers were popular in the field of Natural Language Processing because of their attention mechanisms. The transformers in Computer Vision started gaining popularity when it achieved comparable performances with the state-of-the-arts CNN models [2]. The transformer

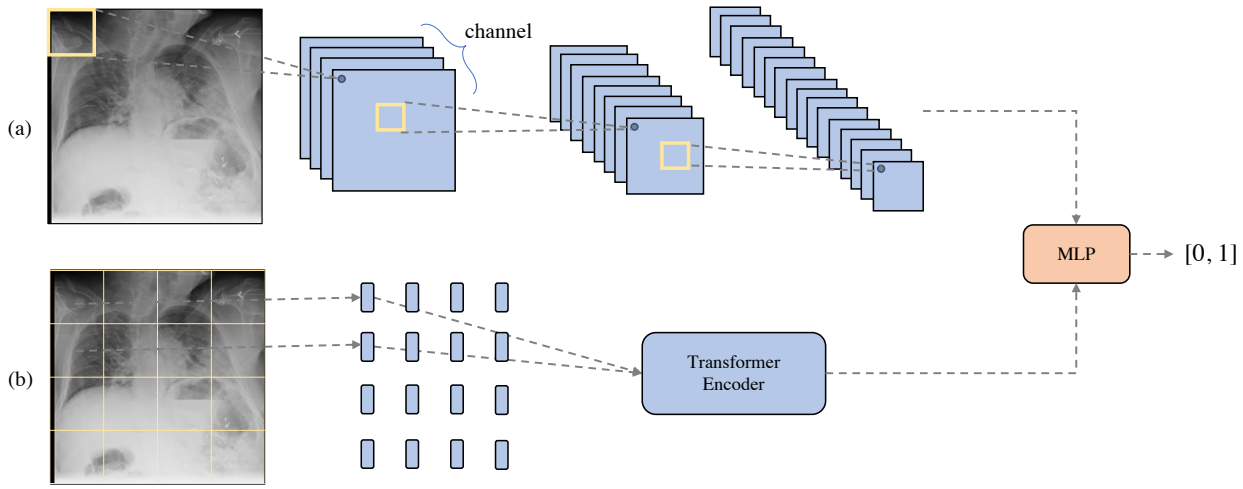


Figure 1. **Convolutional Neural Network vs. Transformer in Binary Classification Task** (a) The general framework of a convolution neural network. (b) The basic framework of a transformer for Image. Note, only a part of the connections between the layers are drawn.

is a deep learning model that employs attention mechanism to weight the relevance of each element of the input data differently. When compared to convolutional neural networks (CNN), Vision Transformer (ViT) offers impressive outcomes while using less computational resources for pre-training [1]. When training on smaller datasets, Vision Transformer (ViT) has a less inductive bias than convolutional neural networks (CNN), resulting in the transformer more relying on data augmentation and model regularization. The ViT model encodes an input picture as a sequence of image patches, similar to the tokens in NLP. When trained on adequate data, ViT outperforms an equivalent state-of-the-art (SOTA) CNN using 4x less Computational resources. In Alexey, D. *et al*'s paper [2], they compared ViT with SOTA CNN models like ResNet on datasets like ImageNet, CIFAR-100. ViT performed better than SOTA CNN models when provided with adequate data to train on.

Behnaz, G. *et al* [3] explores the use of vision transformers for the classification of ultrasound breast images. Two separate datasets on breast US pictures were used in this research. The initial dataset, which is available online, has 780 breast photos from 600 women in the United States, with an average size of image as 500 by 500 pixels. The second dataset contains 163 photos with an average image size of 760 × 570 pixels, divided into two categories: 110 benign masses and 53 malignant masses. The input that was fed into the ViT model was divided into image patches. The Transformer encoder is fed a stream of 1-dimensional patch embeddings, and self-attention modules are used to calculate the relation-based weighted sum of each hidden layer's outputs. As a result, the Transformers may learn global dependencies in the input pictures using this technique. The performance metrics they have used to measure performance were Accuracy and Area under the ROC curve (AUC). They have tried out various architectures of ViT over the dataset of breast US images. The crucial point to note is that the findings for multiple ViT designs were almost same, with roughly 86 percent Acc and 0.95 AUC. The SOTA CNN models like ResNet-50, VGG-16, NASNET were also experimented to delineate the difference in performance measures. The ResNet50 model produced the best results, with an Acc of 85.3 percent and an AUC of 0.95. When they compared their results to the corresponding performance of the SOTA CNNs, they discovered that attention-based ViT models perform similarly to CNN approaches.

Similarly, in Christons, M. *et al*'s paper [7], they explore the use of transformers in classification of medical images by comparing it with CNNs. They ran a series of tests in which they compared vanilla ViTs and CNNs under comparable settings while limiting hyperparameter modification. They chose ResNet50 as the representative of CNN models and DeiT-S with 16 x 16 tokens as the ViT model to provide a fair and interpretable comparison. Some insightful observations were made from their experiments. When trained from start with limited data, the CNNs outperformed ViTs because transformers lacked inductive bias. But surprisingly, for medical tasks, the difference between CNN and ViT performance vanishes when initialized using supervised IMAGENET pretrained weights. On medical imaging tasks, they confirmed that supervised ImageNet pretraining is equally beneficial for ViTs as it is for CNNs. They found that vanilla transformers can consistently substitute CNNs on medical imaging tasks with little effort based on their tests. In smaller medical datasets, ViTs can achieve the

same level of performance as CNNs, but only with the help of transfer learning. Also, the gap between ViT and CNN is likely to widen as the number of samples increases.

In Ze, L. *et al*'s paper [6], they introduce a new type of transformer called Swin Transformer that is capable of addressing the previous concerns in transformer-based models. Large differences in the size of visual elements and the high resolution of pixels in images compared to characters in text are the key hurdles in transformers. They propose a hierarchical Transformer whose representation is calculated with Shifted windows to solve these disparities. By confining self-attention computation to non-overlapping local windows while allowing for cross-window connectivity, the shifted windowing technique improves efficiency. This hierarchical design can simulate at different sizes and has linear computing cost as picture size increases instead of the quadratic computing cost. Swin Transformer's characteristics make it suitable for a wide range of vision applications, including as image classification and dense prediction tasks like object recognition and semantic segmentation. From their experiments, it was observed that the Swin Transformer performed better than ResNet and DeiT-B on ImageNet trained on 22K pretrained models with 1.1% higher accuracy.

3. Method

3.1 Dataset.

To evaluate the performance of the transformer-based models in medical image classification tasks, the randomly sampled subsets from RSNA Pneumonia Challenge in Kaggle are used. In the original RSNA Pneumonia Detection dataset, there are 3 classes, including No Significant Finding (n=8,851), Not Pneumonia / Not Normal (n=11,821) and Pneumonia (n=6,012). Each image has two kinds of labels, one is the disease label and the other is the bounding boxes for the locations of the pneumonia in the image. The dimension of each image is 1024 x 1024 x 1 (in gray color). We only use the disease label in this study for training the models. The bounding boxes for each image are used to visualize the differences between the actual locations of pneumonia and the GradCAM / attention maps predicted from the classification models.

For this study, we firstly divide each class randomly into 10 folds and the following experiments would use different folds for training, validation and testing. There are three experiments: (i) The baseline experiment uses the first two folds as the training set, the third fold as the validation set and the fourth fold as the testing set. Moreover, in this subset, only two classes are used, including No Significant Finding and Pneumonia. All the images for this experiment are resized to 224 x 224, which is similar to the input format for the models we would use. (ii) In the second experiment, the classification task becomes harder. More data is included and the class Not Pneumonia / Not Normal is also kept in the subset for this experiment. That is, the folds 1 to 5 are used for training, the folds 6 and 7 are for validation and the folds 8 and 9 are for testing. (iii) For the third experiment, we would like to know what the performances of these models would be if the input resolution is higher. This is common in medical images because the medical images are always produced in high quality and high resolution. Here, we resize the images to 512 x 512 in order to fit in the GPU memory size. The subset used in this third experiment is the same with the one used in the first experiment. Table 1 shows the overview of these three experiments. In this study, the images with No Significant Finding and Not Pneumonia / Not Normal are grouped into one class, denoted as class 0 and the images with Pneumonia are in the group, denoted as class 1.

Experiment		Subset			Image Size
		Train	Validation	Test	
#1	Class 0	1,771	1,770	1,770	224 x 224
	Class 1	1,204	1,202	1,202	
#2	Class 0	5,609	2,952	2,952	
	Class 1	3,007	1,202	1,202	
#3	Class 0	1,771	1,770	1,770	512 x 512
	Class 1	1,204	1,202	1,202	

Table 1. **Experiment Overview** - #1 and #3 use the same subset. #2 use more data than #1 and #3 and the class (Not Pneumonia / Not Normal) is also included in #2, which makes this subset harder to the model than the other two.

3.2 Model.

We formulate our problem as a binary classification problem $\hat{y} = sigmoid(f(x))$. Given a chest x-ray image (x), the model aims to predict whether it is an image with pneumonia or not ($\hat{y}$). These predictions are further compared to the ground truths (y) for calculating the loss or deriving the metrics mentioned in the following sections. Here, f is the convolution neural network or the transformer-based network. The prediction $\hat{y}$ ranges from 0 to 1.

To investigate the advantages and disadvantages of the transformer-based networks, the convolution neural networks (CNN) are included as part of the baselines. We use the common CNNs with different size of parameters, such as AlexNet, VGG16, InceptionV3 [9], ResNet50 [4], ResNet101 [4] and DenseNet121 [5] to compare with the transformer-based networks, such as ViT [2] and Swin Transformer [6]. Figure 1 illustrates the differences between these two frameworks (CNN vs. Transformer). We can see that the CNN focuses more in the communication between the channel dimension while the transformer learns more about the spatial interactions between the patches. In

ViT, an input image (e.g. chest x-ray in this case) is split into several patches and the patches communicate with each other by using the built-in self-attention module. This helps the model learn the global context of a given image unlike the convolution kernel in a CNN, addressing more on the spatially invariant features. As the time complexity to calculate the interaction between the patches from the whole image is high, the Swin Transformer is designed to calculate the interaction only between the patches in a given window and the window is shifted in the following layers to approximate the interactions between all the patches from the whole image like what ViT does. In this study, we firstly compare the transformers (ViT and Swin Transformer) with the CNNs and then we would discuss the differences between the ViT and Swin Transformer.

3.3 Hyperparameters, Loss Function and Details.

Generally, the hyperparameters for our experiments are set as follows: (i) number of epochs: 10, (ii) batch size: 12 and (iii) Adam optimizer with learning rate 0.0001 and L2 regularization (weight decay 0.00004). Moreover, to prevent model from overfitting, we also apply online data augmentation on the input batch. Random horizontal / vertical flips and random rotation (-15 to +15 degree) are used. The batch size is adjusted to 16 for experiment #2 to boost up the training speed and the number of epochs is set to 20 for experiment #2 and #3 as these two experiment may be harder for the models. Due to the GPU memory limitation, the input image size of the ViT in experiment #3 is changed to 384 x 384. As for the loss function, binary cross entropy loss is applied in this study. All the models in our study are trained based on the pretrained models.

Following each training epoch, the models would be evaluated on the validation set. The best models in terms of the validation loss or the metrics described in the next section would be saved as the checkpoints. After the training process, the best model based on the mean f1 score of class 0 and class 1 would be used to produce the final performances on both the validation and the test sets.

3.4 Evaluation Metrics.

We use the AUC score and the class 1 (Pneumonia) F1 score to evaluate our model performances. To gain more understandings about the model performances besides those two metrics, two major visualization techniques are utilized. For CNN, the GradCAM [8] is used. For the transformer, the built-in self-attention matrix is extracted. These two masks are overlaid with the original chest x-ray and compared with the bounding boxes for the actual locations of the pneumonia.

Model	Model Params	AUC	F1 (Pneumonia)
Alex	61M	0.9696	0.9038
VGG16	123M	0.9813	0.9213
InceptionV3	24M	0.9837	0.9322
ResNet50	26M	0.9806	0.9164
ResNet101	45M	0.9809	**0.9348**
DenseNet121	8M	**0.9850**	0.9339
ViT-Base	86M	0.9770	0.9090
SwinT-T	29M	0.9800	0.9241
SwinT-S	50M	**0.9839**	**0.9315**
SwinT-B	88M	0.9829	0.9297

Table 2. **Experiment 1a** - The performance on the validation set

Model	Model Params	AUC	F1 (Pneumonia)
Alex	61M	0.9682	0.8949
VGG16	123M	0.9789	0.9079
InceptionV3	24M	0.9805	**0.9226**
ResNet50	26M	**0.9825**	0.9192
ResNet101	45M	0.9771	0.9158
DenseNet121	8M	0.9816	0.9177
ViT-Base	86M	0.9715	0.8985
SwinT-T	29M	0.9801	0.9131
SwinT-S	50M	**0.9810**	**0.9178**
SwinT-B	88M	0.9793	0.9158

Table 3. **Experiment 1b** - The performance on the test set

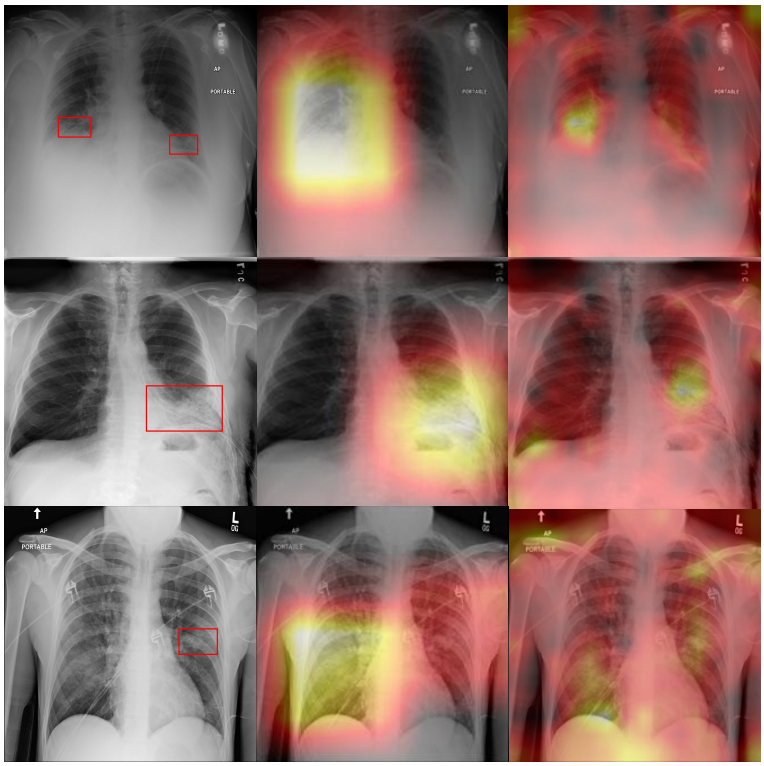
4. Results

4.1 Experiment #1 - Baseline.

Table 2 and Table 3 demonstrate the model performances on the validation set and the test set, respectively. We can see that the performances of the transformers (ViT-Base and SwinT-S) are similar to those of the CNNs. Compared to the performance of the ViT, the swin transformers perform better even with the smallest size of the parameters (SwinT-T). In the experiment #1, the performance gap between the CNNs and the transformers seems to be negligible.

Figure 2 further demonstrate the differences between the CNN and the transformer. Here, the GradCAM is derived from the DenseNet121 and the attention map is drawn from the ViT model. We could see that the DensetNet121 focuses more on one local region in the image while the attention map from the ViT pays attention to several smaller spots. Although both of them seem to be able to capture the lesions of the pneumonia, the higher resolution attention map from the transformer would provide us with more clues on how the model predicts the image as an image with pneumonia.

4.2 Experiment #2 - More Samples & More Difficult Task.



(a) (b) (c)

Figure 2. **Visualization** (a) The Chest Xray image with Pneumonia (Image Size: 224 x 224). (b) The GradCAM of DenseNet121 overlaid on the Chest Xray. (c) The attention of ViT overlaid on the Chest Xray. The bounding boxes in red indicate the location of the pneumonia annotated by the experts.

Model	AUC	F1 (Pneumonia)
ResNet101	**0.9074**	**0.7405**
DenseNet121	0.9055	0.7346
ViT-Base	0.8986	0.7147
SwinT-S	**0.9130**	0.7404
SwinT-B	0.9112	**0.7422**

Table 4. **Experiment 2** - The performance on the test set

Model	AUC	F1 (Pneumonia)
ResNet101	0.9806	0.9163
DenseNet121	**0.9814**	**0.9232**
ViT-Base*	0.9767	0.9097
SwinT-S	0.9744	0.9024

Table 5. **Experiment 3** - The performance on the test set. *The size of the input image for ViT-Base is 384 x 384, which is smaller than that (512 x 512) for the other 3.

In experiment #2, the models are trained with more images and the harder class (Not Pneumonia / Not Normal) is added in class 0. From Table 4, both the AUC score and the F1 score are generally lower than the ones in the experiment #1. Moreover, we could see that the Swin Transformer models perform slightly better than the ResNet101 and DenseNet121 in terms of the AUC score and F1 score. This may indicate that the Swin Transformer is less overfitted with this class-imbalanced subset (number of images of class 0 is nearly twice of that of class 1) and the the transformer may learn more useful features to distinguish the images with pneumonia from those without pneumonia.

4.3 Experiment #3 - Higher Resolution Images.

As the medical images are always stored in the high quality and high resolution format, downsampling the medical images with a large factor may erase the pixels with the lesions inadvertently. In this experiment, we evaluate these

models with higher resolution images (512 x 512 in this case), which may retain more information about the lesions in the images.

Table 5 demonstrates the model performances with higher resolution images. In general, the CNNs achieve higher performances than the transformers. Compared with the results in the experiment #1, the performances of the ResNet101, DenseNet121 and ViT in the experiment #3 are similar or even slightly better. However, we could see that the performance of the SwinT-S slightly decreased. This may be caused by the difference between the input size of this experiment and the input size of the pretrained model of SwinT-S. As the input size is changed with a factor of 2, the attention module in SwinT-S needs to be re-learned, indicating that this module may be not easily transferable between tasks if the input size is different. The performance of the ViT in this experiment seems to be unchanged is mainly because the input size (384 x 384) for the ViT is the same with how the pretrained model of the ViT was trained.

Figure 3 further illustrates more differences between the DenseNet121 and the ViT in the task with higher solution image. We can see that the resolutions of both the Grad-CAM and the attention map get higher than those in the experiment #1 due to the higher resolution input. The Grad-CAM of the DenseNet shows that it attributes its prediction to a local region while the attention map of the ViT demonstrates a spread-out pattern, indicating that several local regions jointly contribute to the prediction.

5. Conclusions and Discussions

In this study, we investigated the advantages and disadvantages of the transformers over the CNNs. For transfer learning, the CNNs are still the first to experiment with. If the size of the input image is similar to the pretrained models of the transformers (e.g. 224 x 224 and 384 x 384), the Swin Transformers could be one of the options. However, with larger size of the image, it may take time for the transformers to re-wire (re-learn) the spatial interactions in the attention module. As a result, the CNNs would be the first choice if the size of the input image is larger. Furthermore, there is an advantage of the transformers over the CNNs. That is, the built-in attention module in the transformers could be easily visualized unlike the situation in CNN, where an additional module (GradCAM in our experiments) needs to be inserted in the model. The resolution of the attention map is generally higher, which would possess more interpretability than the class activation maps in the CNNs.

References

[1] Vision transformers (vit) in image recognition – 2021 guide. https://viso.ai/deep-learning/vision-transformer-vit. Accessed: 2010-09-30. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1, 2, 3

[3] Behnaz Gheflati and Hassan Rivaz. Vision transformer for classification of breast ultrasound images, 2021. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. 3

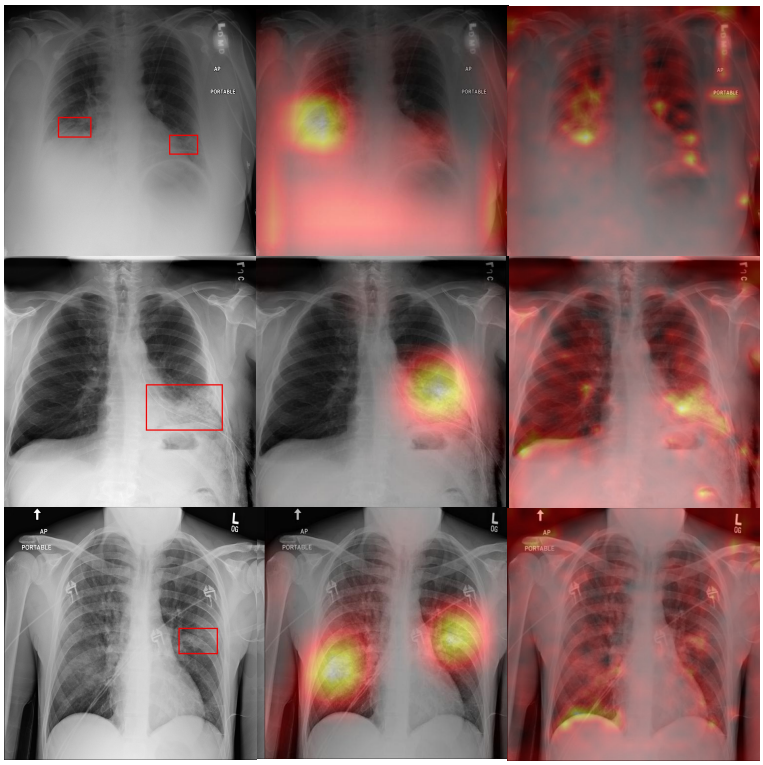
[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 3

[7] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images?, 2021. 2

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. 4

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 3

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1



(a) (b) (c)

Figure 3. **Visualization** (a) The Chest Xray image with Pneumonia (Image Size: 512 x 512). (b) The GradCAM of DenseNet121 overlaid on the Chest Xray. (c) The attention of ViT overlaid on the Chest Xray (Note, the input image for the ViT in the experiment #3 is 384 x 384). The bounding boxes in red indicate the location of the pneumonia annotated by the experts.