# Comparing Transfer learning methods to Classify Pneumonia in Chest X-rays using Grad-CAMs

**Vedant D. Patadia, M.Sc. in Computer Science[1], Manan A. Patel, M.Sc. in Computer Science[2], Dr. Xing Tan, Ph. D[3]**

**[1-3]Lakehead University, Thunder Bay, ON, Canada**

**Abstract**

*The rapid evolution of deep learning applications in the medical field has been a significant assistance to doctors, especially during the recent outbreak of COVID-19 when pneumonia proved to be a major testing ground that had led to vast research in the field. To solve the Chest X-Ray Classification problem, we plan to present a comparative study of three models utilizing a transfer learning method from a dataset from a different domain. As such datasets are ever-evolving, our approach will make use of short-term training on a small dataset with targeted augmentation and activation map visuals to help criticize a model evaluation. This will make it easier for a doctor who has never used such an application before to comprehend the underlying mechanism.*

**Introduction**

For a few years, deep learning architectures have been used extensively in the field of clinical diagnosis. They have proven to be helpful to the fullest, especially during an overload of requests, but at the base, the overall process still needs human supervision, and a lot of suspicions are still present on the implementations. One of the main concerns is how a machine can learn in even a month compared to a professional who has trained for years. With the trends of technological advancements, even doctors are trying to adapt to those solutions. One of the main objectives of the project is to simplify the implementation process for anyone considering it and to give those who use it additional details about the model's approach to the data (through Grad CAM).

Pneumonia essentially refers to a lung infection that primarily affects the lungs' air sacs commonly known as alveoli[1]. According to Johns Hopkins, it is an acute infection in which alveoli fill up with pus or any other liquid[2]. Experts from American Lung Association say that symptoms of pneumonia may include cough, shortness of breath, fever, and shallow breathing[3]. According to the World Health Organization, in 2019, pneumonia alone is responsible for 14% of all deaths of children under 5 years old[4]. While observing Chest X-rays of patients suffering from pneumonia, a cloudy appearance may appear in segments in the lungs or even like patches solely dependent on the airspaces involved inside the lungs, there are a couple of other complications involved in the identification of pneumonia but as referred under [5,6], there might be an increase in lung volume but never an opposite.

The main problem is that no two cases are exactly alike, and even the same cases change over time. Of course, it is outside the scope of our investigation to lay out facts or offer medical diagnosis intuitions, but what we are looking into will shed light on one of the training paradigms known as transfer learning. In particular, the speed in terms of iterations (epochs to be precise) and model inference based on the visual bias (GradCAM[7] Heatmap in our case). For our experiments, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification"[8] by Daniel Kermany, Kang Zhang, and Michael Goldbaum, is a great fit as the dataset is small as it focuses on children below 5 years of age and gives room for grooming it for the transfer learning models.

Transfer learning is a technique for adjusting a particular model that was previously updated on a different dataset[9]. Simply put, the features learnt before are transferred onto the primary set. Most of the time, transferring from a domain that is similar to the primary domain enables the model to be adjusted to the primary dataset. But in our case, obtaining weights for models that were pretrained on a sizable Chest X Ray database is either nearly impossible, comes with conditions, or is not trustworthy enough for short-term tuning. Thus, we go for one of the more widely available datasets, which here is imagenet[10], but one problem here is that the target set for the problem at hand is entirely different from the one we have in imagenet[10]. The basic instinct behind the idea is two parted, one that it will act as a high penalty ground but on second thought pretrained model will contain some lower-level features already optimized. Thus, depending on the random initialization of top layer weights we can see some quick learning saturation.

Looking at the data itself, the scans do have quite a few unnecessary details, like orientation indicators like letter "R" written on them. As such we can also see in Figure 1, on the class imbalance which implies model overfitting is absolute possibility.

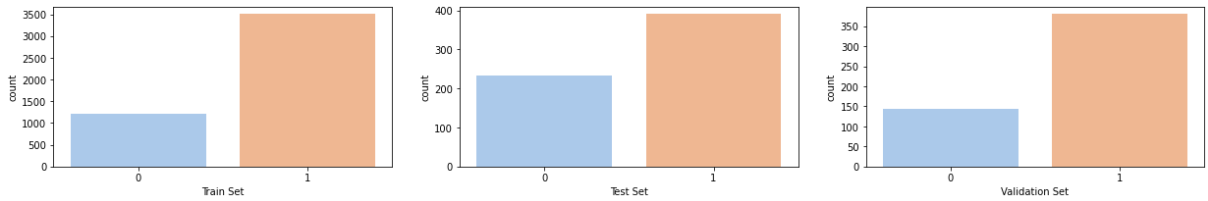


**Figure 1.** Class Imbalance in original dataset without augmentation (**Note:** 0 means Normal and 1 means Pneumonia)

**Related Work**

The research combining Radiology and Deep Learning has been going on well before COVID-19 outbreak and has been quite concrete and focusing on the domain of Deep Learning the method is divided into 4 major structures, Model curation, Transfer Learning, Comparative studies or Surveys and Additive or Adaptations. Here our interests lay in the adaptations and Transfer learning.

**Transfer Learning.** Additionally, many research workers have tried to implement transfer learning on OCT dataset and/or COVID dataset and implementing in Chest X-Ray dataset. For instance, Daniel *et. al.*[11] introduced transfer learning to make a generalized algorithm that classified images for macular degeneration and diabetic retinopathy whilst also distinguishing bacterial and viral pneumonia. On a different account, Samir *et. al.*[12] researched into making transfer learning more stable and quick, by first, investigating impact of classification layer capacity on accuracy, then, focusing in selecting layers to be frozen or not and finally fine-tuning to outperform their best performing model. In [13], authors used Transfer Learning to combine the predictions using a weighted classifier and found an increase of 0.98% in testing accuracy.

**Ensemble Learning.** Another approach of ensemble learning was implemented by authors in [14], where they combine diverse extracted features from chest radiographs to perform classification. Furthermore, Ashitosh *et. al.*[15] performed a detailed comparative study of more than 20 articles based on various factors like data processing techniques, algorithms used, dataset, strengths, weakness, detection of different lung disease like lung cancer, pneumonia, tuberculosis.

**Grad-CAMs.** Ramprasaath *et. al.*[7] proposed "Grad-CAM", a method that uses gradient of any target concept flowing into final convolution layer producing a localization map highlighting vital regions that help in prediction of the image. Dissimilar to any previous visualization technique, Grad-CAM can be used with a wide variety of deep learning models like CNN families, ResNet families and many more.

The rest of the paper accounts to our work as follows: Section 3 addressed a brief description of the methods used in this paper. In Section 4, the results obtained are explained in touch to different parameters along with discussion. This section is followed by Section 5, where we conclude our paper and pilot future enthusiast to carry on more work.

**Proposed Methodology**

For our interest lied dominantly on the activations from the feature extractor, we assumed that the complexity of the top layers could be carried forward from the base model architecture itself. Thus, we used a top layer depth of two with equal units, in our case we used 512 and Fully Connected Neurons.

For the models, we used VGG16, ResNet50 and DenseNet201 models in 2 scenarios with image input of 256*256 and batch size of 16 to carry out comparative analysis for the binary classification of Chest X-Ray (CXR) Images.

1. **Scenario 1:** Considering the original dataset and predicting if the image has pneumonia or not with the plausible justification from the model Grad CAM.
2. **Scenario 2:** Applying a series of augmentations to the training set and the making predicting if the image has pneumonia or not with the plausible justification from the model Grad CAM. We performed more augmentation on the normal images (all augmentation from pneumonia images plus Horizontal Flip) to approximately balance the class imbalance (Figure 2). The list of augmentations that are performed on the normal images are as follows:
    a. *Rotation:* a random rotation ranging from 20° in clockwise as well as anti-clockwise direction.
    b. *Horizonal Flip:* making a horizontal flip of the image.
    c. *Contrast:* changing contrast with a factor range of 0.1.
    d. *Brightness:* changing brightness with a factor range of 0.1.
    e. *Hue & Saturation:* changing hue and saturation for an image with a factor range of 10 and 20 respectively.

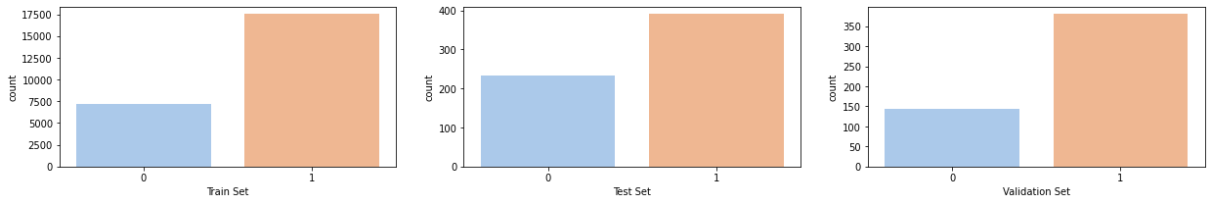


**Figure 2.** Class Imbalance in modified dataset with augmentation

The first scenario is quite straight forward but in the second one we tried to address the class imbalance even if little bit. Furthermore, we added dropouts to the top layers to decrease overfitting, and step learning rate scheduler to decrease Learning Rate every 10 epoch for a better curve and less reliance on random weight initialization.
Typically, in transfer learning the model is divided into two parts, one is the feature extractor, and the top called the classifier, where the feature extractor is kept frozen from training to keep the optimized learnt features intact and thus not disturbed by the random initializer and high learning rate.
This is for a general scenario, where high level features of both datasets involved in transfer learning is similar at worst, thus in our case we assumed that, without enabling training of feature extractor (even if partially) a high-level change in domain will be difficult to explain. Thus, we introduced a new variable to the mix that we call saturation threshold. whose value corresponds to the layer at which classifier is bound to be saturated or close to saturation, thus after which we can enable training of the feature extractor.
In summary, we trained the model while keeping the feature extractor frozen, which we call fine tuning phase. After reaching the threshold we enable training of feature extractor, thus going through final tuning of the model.
The value of threshold in our case was around 10 or less, as it was a tendency of such high-capacity models to overfit on a binary classification problem of grayscale images. Also, in our experiments we observed that if threshold is kept too low then the model erratically shoots around its weights thus leaving a generalization to random coincidence. But if the threshold is too high then we will have to forcefully train the model more so as the smaller loss would tune the model to our needs.

**Experimental Results and Discussion**

We devised 6 experiments in total under the three models with the same threshold to support comparisons (Table 1), with which we observed that scenario two is at least slightly better than the scenario one in ResNet and DenseNet. But with VGG16 it was quite evident that augmentations play a role in improving performance.

**Table 1.** Model result comparison

| | **VGG16** | | **Resnet50v2** | | **Densenet201** | |
|---|---|---|---|---|---|---|
| | **Scenario 1** | **Scenario 2** | **Scenario 1** | **Scenario 2** | **Scenario 1** | **Scenario 2** |
| **Accuracy** | 0.71 | 0.82 | 0.73 | 0.74 | 0.73 | 0.75 |
| **Precision** | 0.62 | 0.76 | 0.64 | 0.66 | 0.64 | 0.67 |
| **Recall** | 0.84 | 0.87 | 0.84 | 0.82 | 0.84 | 0.84 |
| **F1-Score** | 0.60 | 0.78 | 0.64 | 0.66 | 0.62 | 0.67 |

Here we assumed that the highest probability of this happening is due to the model depth and residual nature of ResNet and DenseNet, that they either failed on final tuning or the classifier capacity was not enough. This of course can be looked upon in the training loss curves.
The loss curves on other hand show a different picture at the front showing VGG and DenseNet clearly overfitting and even starting at quite a low magnitude of loss but looking closely and expanding on the axes we are able to see a trend and also the effect of random weight initializer at play.
The overall loss curve of VGG16 in Figure 3, is quite similar in both scenarios while the one with augmentations starts quite low while showing signs of underfitting (which is good). Also looking at around $10^{th}$ epoch we can see a general descent, and quite a smooth one.

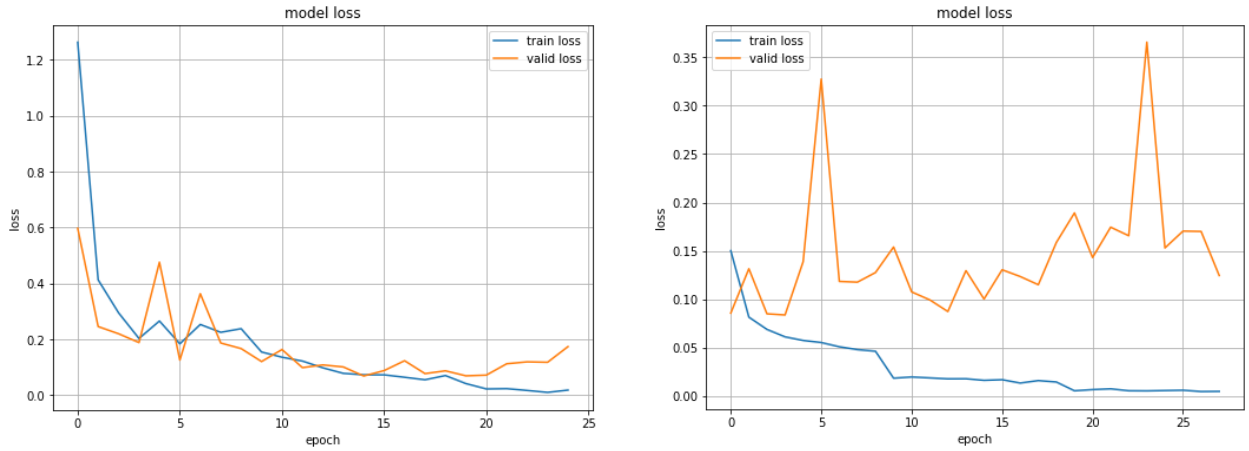


**Figure 3.** VGG16 Model History (Left) and VGG16 Augmented Model History (Right)

The case of ResNet in Figure 4 is almost similar to VGG but looking aside the validation loss we can see that the training is quite stable on the left without augmentation, whereas we can see a general curve with augmentations. Which shows us that in this case augmentation played a role in adding penalty to the regime. Although looking closely the model still tends to overfit in either case.

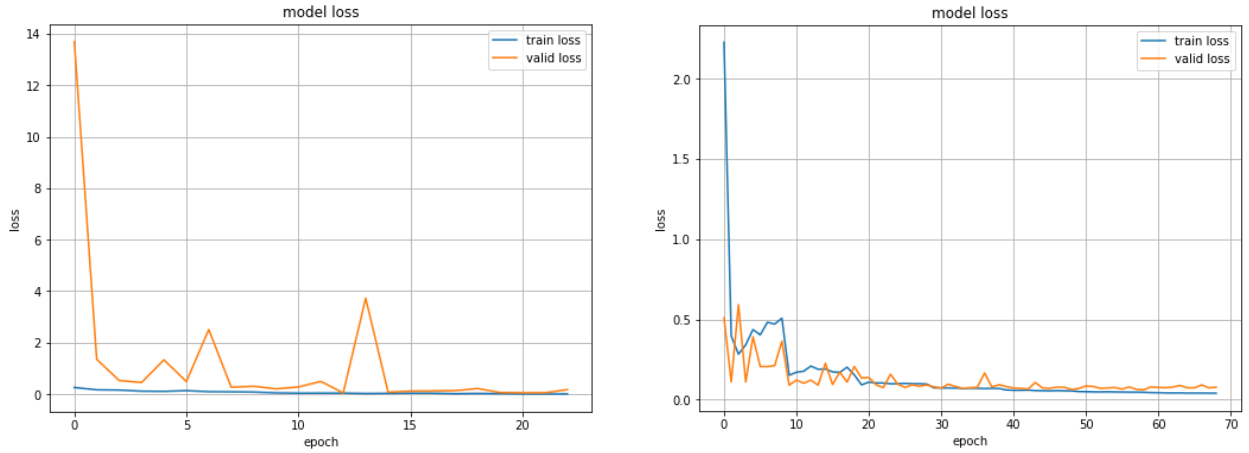


**Figure 4.** Resnet50v2 Model History (Left) and Resnet50v2 Augmented Model History (Right)

Furthermore, in the curve without augmentations show little to no fluctuations around epoch 10 on training curve but just a slight bump and then a bigger one on validation curve, which might just be because of learning rate change or random deviation, but looking back a similar trend is also shown in VGG16. But moving onto the one with augmentations, we can see a clear drop before the threshold (10) epoch, and which shows an overall improvement due to training of feature extractor.

DenseNet on another end shows a different trend on itself, without augmentations, the curve is as smooth with small but minor bumps on only validation curve similar to ResNet but no variations around threshold as shown in Figure 5.

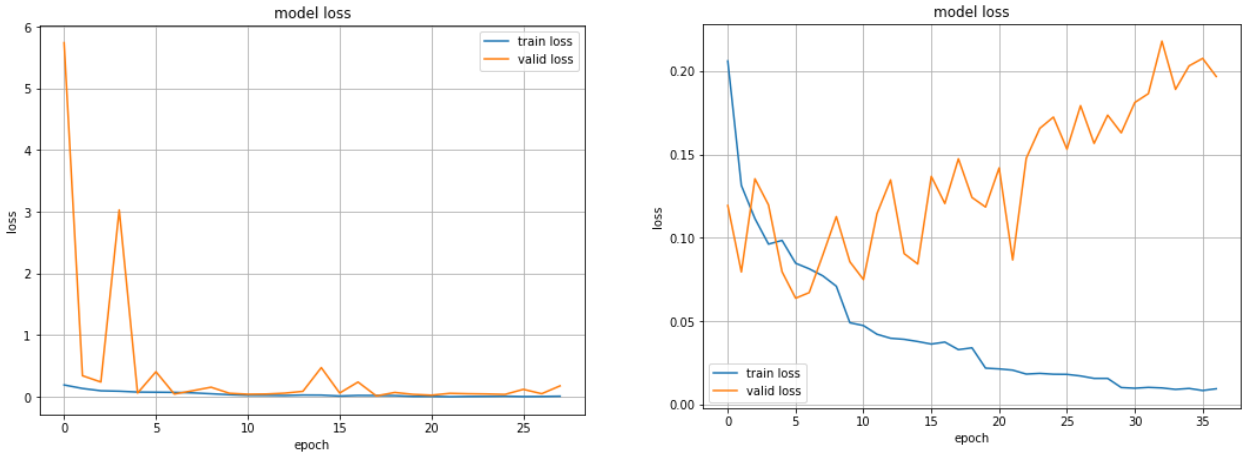


**Figure 5.** Densenet201 Model History (Left) and Densenet201Augmented Model History (Right)

In Addition to that, the curve with augmentation is quite erratic on validation with no definitive bump after threshold epoch but a slight general descent is on the training curve after threshold. This is no clear indication of what to assume but a little more digging around is needed.

Now translating our curves onto the generated Grad-CAM, we see a difference in trend in the case of DenseNet, but one major aspect to keep in mind, our discussion will majorly depend on the mid row. Also, for reference, in each bounding box of Figure 6, 7 & 8 showing Grad-CAMs, the very first column shows Normal images and the rest two are listed to be diagnosed with Pneumonia.

In VGG16, we can see the model is looking into generalization after it is given the dataset with augmentations which is quite evident in the superimposed images (bottom row) of the very last prediction (first from right), showing the interest areas around lungs. All in a very general improvement is shown already as looking at the results without augmentations are quite plain and the ones with augmentations show a slight structure.

As discussed earlier the interest areas are not something we can comment upon as the disease spread is although available to public reading, but the diagnostic knowledge needed to comment upon is not our domain. But looking at the Grad-CAMs in ResNet, without augmentations the model is quite simply looking near the diaphragm and slightly above it while the other area seems quite irrelevant or simply deductions. But in the case of augmented data the model starts looking into the general structure of the human body and the interest areas are getting more generalized.

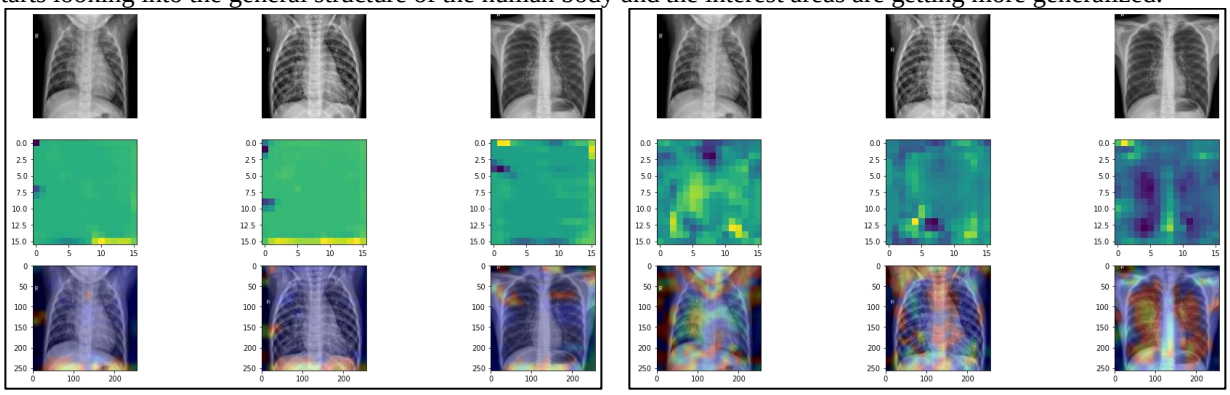

**Figure 6.** Grad-CAM VGG16 without augmentation (Left) and VGG16 with augmentation (Right)

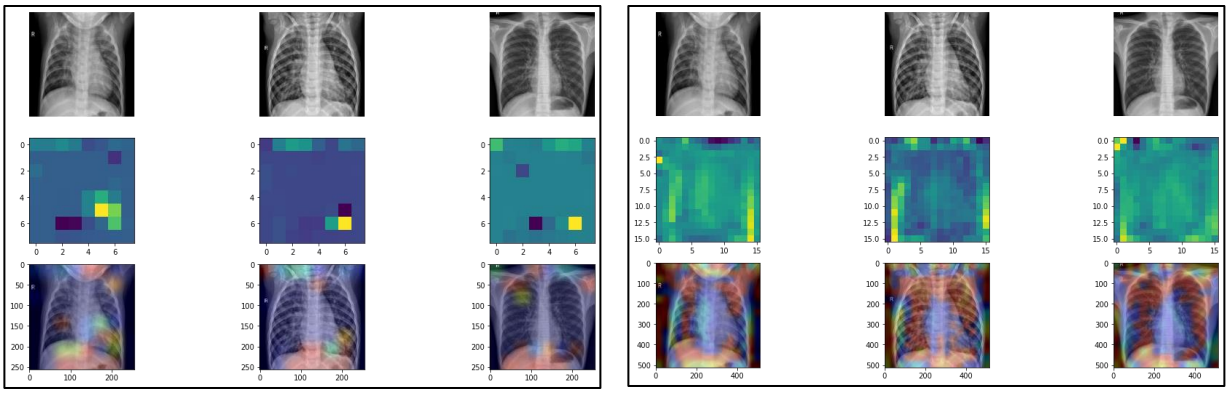

**Figure 7.** Grad-CAM ResNet50v2 without augmentation (Left) and ResNet50v2 with augmentation (Right)

Although in this case the effect of mid separation is not quite evident in heatmaps (mid row) but looking at the superimposed images (bottom row), it does seem that the model is targeting the lung areas for activations.

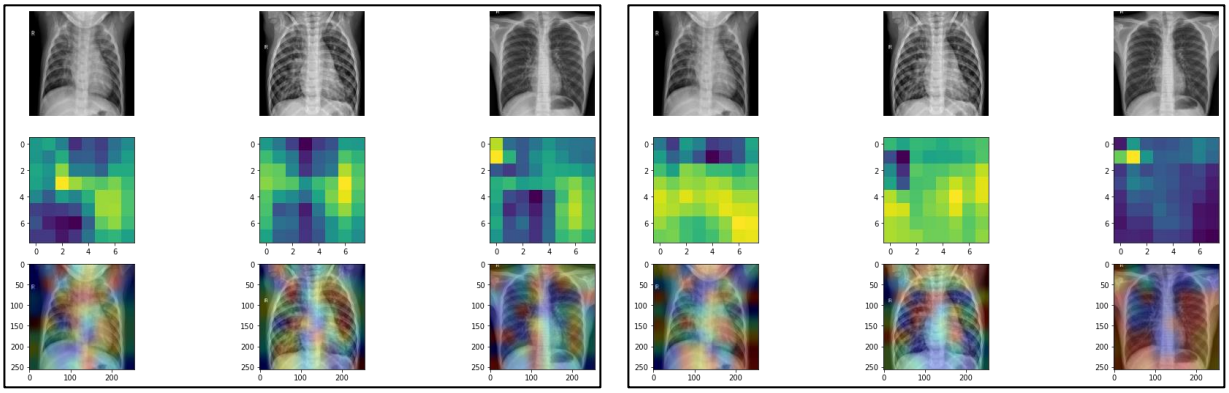

**Figure 8.** Grad-CAM DenseNet201 without augmentation (Left) and DenseNet201 with augmentation (Right)

The Grad-CAM visuals in DenseNet on other hand seem quite indifferent on a glance, but inferring from Table 1, and Figure 5 that do seem to fare correctly. Either it be with or without augmentations, the training loss is already quite low which we can safely assume that due to the depth of the model structure and already known weights the learning process is quite slow or even stagnant. Although we can see structures around in the heatmaps but translating them on the X Rays, at most we can say that without augmentations there is an evident structure on lungs along with separation of spine. But the same doesn't transfer to the one with augmentations, which can be due to low penalty leading to less precision on short term training and incorrect threshold setup, the prior is already evident after observing the VGG16 and ResNet50, also can be confirmed by testing on DenseNet101. The latter on the other hand can be tested by training on different lengths but out of our scope.

**Conclusions and Future Scope**

From the base we can see that the threshold we set on final tuning plays an important role in how the model slides back into a fit. Of course, our interpretation on X Rays as a medical reference is trivial but in conclusion to the training process, we can clearly see that without augmentation it would take more work, time and computational power to get a good fit however be the case of the top layers. Furthermore, we were able to see clear results on Grad-CAM that augmentations played a role in inducing the required penalty required to improve a model further, this is most dominant in VGG16 followed by ResNet and then DenseNet. The same is conclusive with the overall score with VGG16 being the best model thus.

Our results are quite limited to binary classification and simplistic top layer to decode vectors from feature extractor, also the efficacy of the results is relied upon random weight initialization. The future research can be done on, adding more classes like, ChestX-ray8[16], COVID-19 Pneumonia, Viral Pneumonia and Bacterial Pneumonia, and a more sophisticated top layer while improving efficacy with more reliable weight initializer may help build better training

curves. Finally, one of the variables we didn't experiment with was increasing the input size, which we stuck with 256 as it's closer to the initial input size to imagenet[10].

## References


1. American Lung Association. Learn About Pneumonia | American Lung Association [Internet]. Lung.org. American Lung Association; 2019. Available from: https://www.lung.org/lung-health-diseases/lung-disease-lookup/pneumonia/learn-about-pneumonia
2. John Hopkins Medicine. Pneumonia [Internet]. John Hopkins Medicine. 2019. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/pneumonia
3. American Lung Association. Pneumonia symptoms and diagnosis | American Lung Association [Internet]. www.lung.org. American Lung Association; 2020. Available from: https://www.lung.org/lung-health-diseases/lung-disease-lookup/pneumonia/symptoms-and-diagnosis
4. World Health Organization. Pneumonia [Internet]. Who.int. World Health Organization: WHO; 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/pneumonia
5. X-ray Atlas: Chest X-ray | GLOWM [Internet]. www.glowm.com. Available from: https://www.glowm.com/atlas-page/atlasid/chestXray.html
6. Weerakkody Y. Reticulonodular interstitial pattern | Radiology Reference Article | Radiopaedia.org [Internet]. Radiopaedia. [cited 2023 Mar 21]. Available from: https://radiopaedia.org/articles/reticulonodular-interstitial-pattern
7. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-{CAM}: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision [Internet]. 2019 Oct;128(2). Available from: https://doi.org/10.1007%2Fs11263-019-01228-7
8. Kermany D, Zhang K, Goldbaum M. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. datamendeleycom [Internet]. 2018 Jun 1;2. Available from: https://data.mendeley.com/datasets/rscbjbr9sj/2
9. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big Data. 2016 May 28;3(1).
10. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision [Internet]. 2015 Apr 11;115(3):211–52. Available from: https://hci.stanford.edu/publications/2015/scenegraphs/imagenet-challenge.pdf
11. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell [Internet]. 2018 Feb;172(5):1122-1131.e9. Available from: https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5
12. Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data. 2019 Dec;6(1).
13. Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW. Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning. Diagnostics. 2020 Jun 19;10(6):417.
14. Masud M, Bairagi AK, Nahid AA, Sikder N, Rubaiee S, Ahmed A, et al. A Pneumonia Diagnosis Scheme Based on Hybrid Features Extracted from Chest Radiographs Using an Ensemble Learning Algorithm. Singh D, editor. Journal of Healthcare Engineering. 2021 Feb 25;2021:1–11.
15. Tilve A, Nayak S, Vernekar S, Turi D, Shetgaonkar PR, Aswale S. Pneumonia Detection Using Deep Learning Approaches [Internet]. IEEE Xplore. 2020 [cited 2021 Nov 19]. p. 1–8. Available from: https://ieeexplore.ieee.org/abstract/document/9077899
16. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. 2017 Jul; Available from: https://arxiv.org/abs/1705.02315