Data Augmentation for Aiding Pneumonia Classification

Matthew Carter[††], Ivan Chuprinov[†], Chris Drazic[‡], Alex Gavin[§] and Raleigh Hansen[¶]
Department of Computer Science, Western Washington University
Bellingham, Washington USA
Email: [††]carter36@wwu.edu, [†]chuprii@wwu.edu, [‡]drazicc@wwu.edu, [§]gavina2@wwu.edu, [¶]hansen92@wwu.edu

I. Overview

The identification of pneumonia in chest x-rays is frequently misclassified by experts, and in low resource countries, there is often a lack of trained radiologists to perform such identification [1]. Computer aided diagnosis systems (CAD) seek to solve this issue by automating some or all of the identification process. Our aim was to further this field of research by focusing solely on the effect of data augmentations on the performance of pneumonia CAD systems using a widely available, expert-verified dataset. Initially, we had sought to explore the performance impact of data augmentations on several different classifiers. However, due to time constraints, we narrowed our focus to only the task of classifying chest x-rays of lungs which either exhibit symptoms of pneumonia or do not exhibit symptoms.

II. Background

As we have seen in class, convolutional neural networks (CNNs) are widely used in image classification tasks. While there exist many CNN architectures which have shown promising results in medical classification tasks at large, one especially effective architecture is the ResNet series [1], [2]. A common theme in the post-AlexNet era of CNNs is that of "going deeper" or adding more layers to achieve more expressive models [3], [4], [5]. However, as the depth of standard neural networks increases, so too does saturation of accuracy [4]. The aim of ResNet is to mitigate this decrease in performance by using what are called residual blocks. To do so, residual blocks employ an identity mapping, referred to as a skip connection, which simply feeds the pre-residual block input to a layer "further along" in the architecture. The creators of ResNet were able to see sizable increases in accuracy as well as a marked decrease in trainable parameters when compared to models that were then state of the art [4].

Upon a more detailed inspection of the dataset for the task of pneumonia classification, we noticed that the data skews heavily towards one class; that is, there are significantly more pneumonia images than there are normal, non-pneumonia images. This class imbalance property occurs frequently in medical image sets, so researchers often employ data augmentation to counteract the negative effects such an imbalance can have on performance [6]. As we will demonstrate in the Prior Work section, data augmentations have the added benefit of promoting better model generalization.

When utilizing imbalanced datasets, one must consider how imbalances may affect performance. In addition to accuracy, there exist other, more expressive performance metrics such as precision, recall, F1 score, and the precision-recall curve which are well-suited for evaluating performance given skewed data [7]. The first three scores are defined as follows:

$$Precision = \frac{T_p}{T_p + F_p} \tag{1}$$

$$Recall = \frac{T_p}{T_p + F_n} \tag{2}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

Where T_p is defined by true positive and F_p is defined by false positive.

As an example, suppose there exists a binary classifier which makes predictions on data labeled from the set $\{A, \neg A\}$. Precision effectively determines "how good" this classifier is at not predicting a positive (A) having seen a negative ($\neg A$). Recall, on the other hand, determines "how good" the classifier is at predicting positive (A) having seen a positive (A), something that is especially important in medical applications. In the context of CAD systems for pneumonia classification, high recall ensures that actual cases of pneumonia don't "slip under the radar" whereas high precision enables experts to have high confidence that x-rays classified as pneumonia actually exhibit symptoms of pneumonia. To sum up the information of both of these scores, the F1 Score takes the harmonic mean of both the recall and precision. As noted in [8], there are different methods to calculate f-scores, some of which exhibit inherent biases possibly affecting performance. Considering this, we note that scikit-learn's metric implementation used does not use the recommended method and, therefore, may exhibit biases of approximately -5% or +12%.

To further avoid the inherent bias caused by class imbalance, models may be formally evaluated using metrics which do not rely on having a balanced dataset. In cases where a class imbalance is present, it is common to rely heavily on a precision-recall curve (PRC plot). This provides a more accurate measure of future predictions some model may make given an imbalance within its training data. PRC plots show the overlap in predictions for true positive and true negative

data points for all given concentrations of true positives. Thus, these plots provide a way to compare model performance without concern for skewed results caused by some class imbalance that are seen in other performance metrics [7]. PRC plots alongside confusion matrices are often provided to give a clear look at predictions made by each model. This reflects the work done to mitigate the impact of imbalance within the data.

III. Prior Work

Much of the recent work in pneumonia CAD systems has focused on improving the performance of existing models. One such work has leveraged such models as VGG19, MobileNet_V2, and ResNet50 to demonstrate the efficacy of end-to-end deep learning CAD systems [2]. This research showed that ResNet50 and MobileNet_V2 performed much better than other models with accuracy scores above 96% and sensitivity scores above 94% when classifying pneumonia in X-Ray images [2]. Some data augmentations were used to increase the balance in the dataset but were not the main focus of the work. In [9], researchers took this one step further by leveraging a modified AlexNet (MAN) coupled with handcrafted features for the same task. Their approach, which was also used in classifying malignant lung cancer in CT scans, demonstrated precision, recall, F1, and accuracy scores each in the 96.5-96.99% range [9].

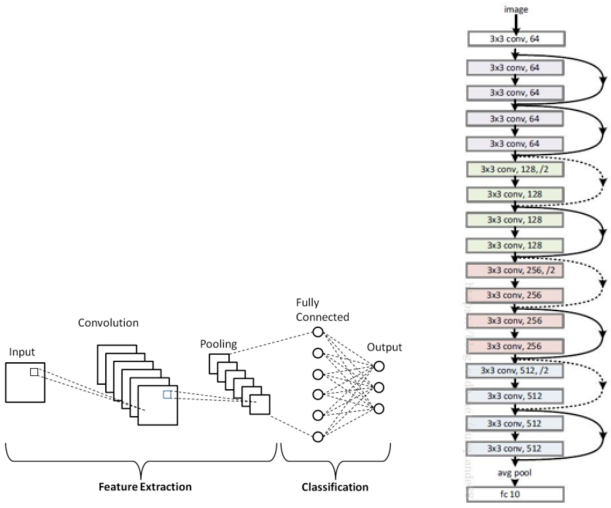
Continuing to leverage existing architectures, Rahman et al. [1] demonstrated strong boost in performance when using transfer learning on such models as ResNet18 and DenseNet. This research showed that DenseNet when pretrained on the ImageNet dataset outperformed its peers, especially in training multi-class classifiers which identify bacterial versus viral pneumonia [1]. Note that this research used scaling, rotation, and translation to artificially increase the size of their dataset. However, the authors did not specify how many images from the unaltered dataset they used to perform the augmentations nor did they specify what combinations of augmentations were used.

In another work leveraging data augmentations for pneumonia classification, researchers performed experiments on a dataset of top-down chest x-rays exhibiting pneumonia or no pneumonia [10]. These experiments either increased the contrast, increased the brightness, or performed color scheme expansion on their dataset. The authors found that increased contrast resulted in the largest performance boost when using ResNet9 [10].

A more general work on data augmentations, Shorten and Khoshgoftaar [6], explored the effect of different augmentation styles such as photometric and geometric approaches. They found that the best performing augmentations in the papers surveyed were brightness and contrast, along with horizontal flipping, vertical flipping, and rotation. Increases in accuracy ranged from 1.5-2.5% from the papers surveyed [6]. This research demonstrates potential utility of such data augmentations for the task of pneumonia classification.

IV. Methodology

To aid in the comparison of our results, we decided to train both a baseline model using a simple CNN as well ResNet18, an established model in this field of research. Our goal was to use the simple CNN to further verify previous results in pneumonia classification which indicated that deeper, more expressive models such as ResNet18 perform better than a simple CNN. Furthermore, we aimed to demonstrate the effect of data augmentations on the performance of ResNet18.



(a) Simple CNN (b) ResNet18

Fig. 1: Visual description of our models

For the design of the simple CNN, as shown in figure 1a, we opted for a straight feed forward architecture, which consists of a single convolutional layer followed by an activation function then downsampling the input through a max pooling layer. Finally, the input is flattened out through a fully connected linear layer and we arrive at a classification. To utilize an established architecture within this domain, we chose to use the ResNet18 architecture shown in figure 1b, since it has been shown to demonstrate better results than a standard CNN and has less trainable parameters than larger ResNet architectures [4]. Each of these architectures requires an input size of $[224 \times 224 \times 1]$, requiring resizing and gray-scaling of images before training.

We utilized the PyTorch toolkit for the implementation of both CNNs and leveraged PyTorch's pre-built implementation of ResNet18 for the experiments. Additionally, we used PyTorch's torchvision library to perform our chosen data augmentations using the available transform functions. Finally, we used the scikit-learn toolkit coupled with numpy to calculate model metrics, which were logged using the python library and web app, Weights and Biases (wandb).

In order to train and test our models, we made use of the WWU research computing cluster. We specifically used the training script provided to us for hyperparameter tuning and augmentation selection as well as final model training. Each training model coordinated the logging and saving of metrics,

final model parameters, and other data with wandb. The testing script, on the other hand, coordinated the loading, evaluation, and logging of final test metrics on all trained models using default data (no augmentations).

Brightness	Brightness + Contrast
Contrast	Brightness + Rotation
Rotation	Brightness + Contrast + Rotation

Fig. 2. Table of Augmentations.

Using observations from previous research, we determined that combinations of brightness increases, contrast increases, and rotations were promising for pneumonia classification. These augmentations are listed in figure 2. To isolate the effects of distinct data augmentations as effectively as possible, we deviated from the norm of using data augmentations to increase the size of the dataset. We instead opted to apply the chosen augmentations to the unaltered data on every run. Note that we apply the augmentations before resizing so as to minimize any potential information loss since the images are of varying shapes. As specified in figure 2, such augmentations were performed randomly within predetermined boundaries.

V. Experimental Results

Having chosen an expert-validated dataset of 5863 labeled chest x-rays presorted into train, dev, and test sets, we chose to shuffle and re-sort the entire dataset into new train, dev, and test sets in the ratio 60/20/20 respectively [11]. As mentioned previously, this dataset contains significantly more chest x-rays of pneumonia than not. In evaluating the performance of our chosen augmentations, we consider this property.

Given the inherent variability of both the data augmentations applied as well as the model parameters, we trained each model and averaged performance over five runs. This number of runs was chosen as a trade off between increasing sample size and remaining cognisant of time constraints. We average over performance metrics discussed earlier to perform evaluation. However, we exclude the PRC plot evaluation during the bulk of training. The data used in creating these PRC plots require much more memory than is required to compute precision, recall, and F1 scores. As both time and GPU memory are limited resources, we are limited to using PRC plots only for the final evaluation and testing of our models. As discussed previously, PRC plots provide a valid means of comparing performance across models trained on a dataset with significant class imbalance, as their results are independent of an imbalance present. With PRC plots being unavailable for during the hyperparameter tuning process, we seek to strike a balance with high values of precision, recall, and F1 score when training models.

With such metrics in mind, we separated our work into a few distinct stages, each executed consecutively. First, after implementing and testing our models, we performed hyperparameter tests on both models with no data augmentations to guide training when using data augmentations. Next, we trained new ResNet18 models on the chosen data augmentations (including no augmentations) each for five runs using the best performing hyperparameters. Finally, we tested all trained models using no data augmentations on the test set in order to determine final results.

The desirable hyperparameters for the baseline consisted of a ReLU activation, learning rate of 0.001, minibatch size of 8, and the Adam optimizer. These hyperparameters were found through the elimination of under performing parameters and large scale testing of well-performing parameters. Similarly for ResNet, the hyperparameters were tuned in the same manner and were all the same except for the use of the AdaGrad optimizer instead of Adam.

The evaluation metrics used during training were accuracy, precision, recall, and F1 scores. These metrics were used to provide a general view of how well a model and its hyperparameters were performing as PRC plots were too computationally expensive during training. However, during the final testing of our trained models, we utilized PRC plots and confusion matrices in order to better visualize the performances of our trained models. The results for both the baseline and ResNet models can be captured in the following confusion matrices:

		Predicted Positive	Predicted Negative
Actual	Positive	807	47
	Negative	37	249

Fig. 3. Baseline Confusion Matrix.

		Predicted Positive	Predicted Negative
Actual	Positive	827	27
	Negative	31	285

Fig. 4. ResNet Confusion Matrix.

While the results on the test set across ResNet models trained with augmentations can be visualized in the following PRC graphs.

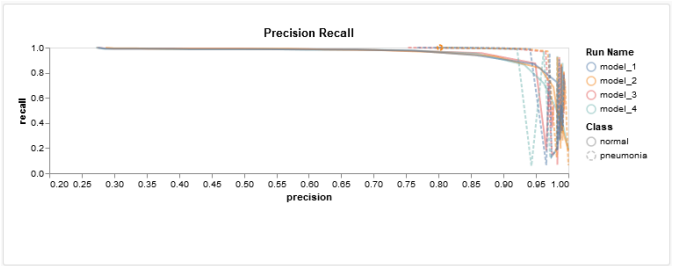


Fig. 5. No augmentations.

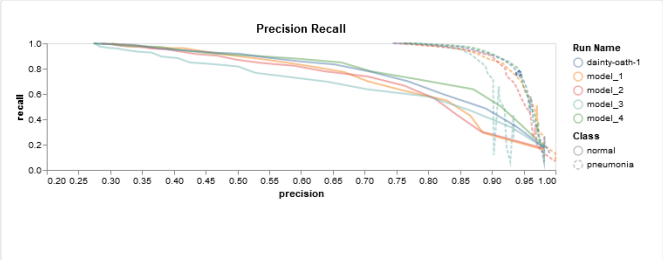


Fig. 6. Brightness boost.

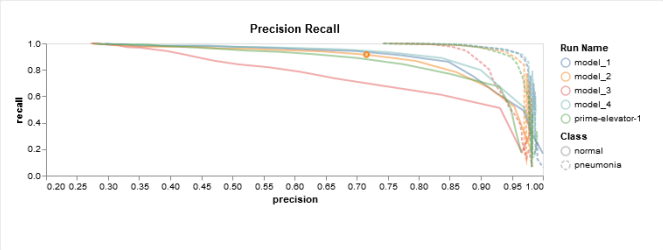


Fig. 7. Contrast boost.

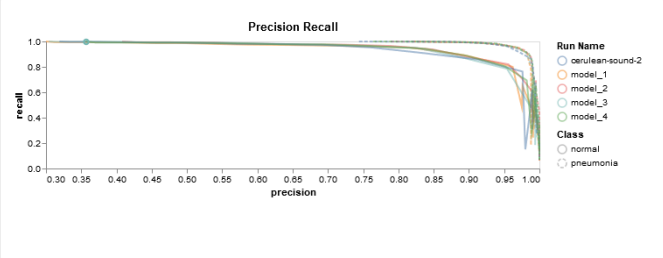


Fig 8. Random rotations.

These PRC graphs show that the ResNet models trained with no augmentations and ResNet models trained with only rotation augmentations provide the best results. The other models trained on combinations of augmentations did not perform nearly as well. The only exception to this being the combination of contrast and rotation which showed slightly better results than just contrast alone.

VI. Conclusions and Future Work

According to the results of our experiments, ResNet18 continues to outperform simpler architectures, verifying previous work. Additionally, results show that data augmentation when used in a way that does not to increase the size of a dataset generally does not improve performance, especially when augmentations are combined. Randomly applying rotation to images did smooth the PR Curve for pneumonia prediction compared to applying no augmentations. However, the overall results of rotations indicate that it, like other augmentations, does not improve performance as a whole. Regarding combinations of data augmentations, we suspect that our implementation may have exacerbated the decreases in performance demonstrated as it is possible for multiple augmentations to be applied to the same image.

Considering the results of our work and previous research, there exist several areas where this work can be improved and extended upon. First, changing the used method of applying the data augmentations to ensure at most one augmentation is applied to an image would be a good place to start, as our results inherently rely on how the augmentations are applied. Furthermore, since the images in the dataset were of a variety of shapes, the method and location of resizing the images in the sequence of applied transformations could have a large effect on results. Finally, parameters for the employed augmentations may warrant investigation. Our review of previous work identified promising augmentation types but not necessarily the parameters for which these augmentations function best. Considering the variety of unknowns regarding our work, we believe that investigating the mentioned areas of future work may be more beneficial than simply investigating the same problem with a different architecture.

We continue to support further research into the current viability of CAD systems using deep learning. With more research into deep learning methods for CAD systems occurring and the continued creation of datasets of successively higher quality, this field of research holds great potential for improving access to targeted healthcare in under-served communities.

References

[1] T. Rahman, M. E. H. Chowdhury, A. Khandakar, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Kadir, and S. Kashem, "Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, May 2020.

[2] K. E. Asnaoui, Y. Chawki, and A. Idri, "Automated methods for detection and classification pneumonia based on x-ray images using deep learning," 2020.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, p. 60, 2019.

[7] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, 2015.

[8] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *SIGKDD Explor. Newsl.*, vol. 12, no. 1, p. 49–57, Nov. 2010.

[9] A. Bhandary, G. A. Prabhu, V. Rajinikanth, K. P. Thanaraj, S. C. Satapathy, D. E. Robbins, C. Shasky, Y. Zhang, J. R. Tavares, and N. S. M. Raja, "Deep-learning framework to detect lung abnormality – a study with chest x-ray and lung ct scan images," *Pattern Recognition Letters*, vol. 129, pp. 271 – 278, 2020.

[10] C. Saul, D. Urey, and C. Taktakoglu, "Early diagnosis of pneumonia with deep learning," *CoRR*, 2019.

[11] D. Kermany, K. Zhang, and M. H. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," in *Mendeley Data*, vol. 2, 2018. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj/2