Investigation of Artificially Distorted Chest X-Ray Images in CNN Based Pneumonia Detection

Vikram Aikat VAIKAT@LIVE.UNC.EDU
Ryan Armstrong RYANAA@LIVE.UNC.EDU
Rohan Arora RARORA9@LIVE.UNC.EDU
Sam Nielsen SNIELS@LIVE.UNC.EDU

Abstract

Previous work has has applied deep learning methods to classify pneumonia in chest x-rays. These methods rely on training and testing on quality x-ray images. However, approximately 5% of chest x-rays experience an augmentation that requires image retake for diagnosis, doubling patient radiation exposure. Using a convolutional neural network (CNN) architecture based on state of the art pneumonia detectors, we quantify the effects of clinically relevant augmentations to investigate if CNNs can be used to limit image retake. Notably, we found a CNN only trained on clean images is not robust to imaging issues, however, it retains accuracy to a larger degree in augmentations that humans are poor at reading (such as blurring effects) than in augmentations humans are better at reading (such as underexposure), indicating possible motivation to further explore the use of CNNs to detect pneumonia in these areas.

1. Introduction

Precise, clinical detection of Pneumonia requires chest x-ray imaging (Wootton & Feldman, 2014). However, anatomy cutoff, artifacts, patient movements, and machine problems cause inadequate image quality for diagnosis in $\sim$ 5% of chest examinations (Lin et al., 2016). These occurrences lead to image retake and therefore double the patient's exposure to carcinogenic radiation (Berkhout, 2015).

Thus, this study investigates the use of a convolutional neural network (CNN) image classifier to identify pneumonia in poor quality chest X-ray images with the goal of decreasing image retake rates, thereby reducing overall patient radiation exposure. More specifically, the study does not focus on developing the most accurate CNN for detecting Pneumonia, but rater explores the robustness of a generic Pneumonia classifying CNN to image degradation by quantify the effects of different distortions on these models. Ultimately, the project looks to demonstrate whether or not classically rejected x-ray images are actually diagnosable with the assistance of appropriate CNNs and therefore do not require a second exposure.

1.1. Literature Review

The application of deep learning methods to the detection of pneumonia in chest x-rays is not new. Notably, the the Stanford Machine Learning Group developed the CheXNet algorithm which detects pneumonia at a level exceeding practicing radiologists (Rajpurkar et al., 2017). The algorithm uses a 121-layer Dense Convolutional Network (DenseNet) structure (Huang et al., 2017) that was pretrained on the ImageNet (Deng et al., 2009) data-set. The model was then applied to the largest data-set of annotated chest x-ray images, ChestX-ray14 (Wang et al., 2017).

Importantly though, the ChestX-ray14 data only includes non-rejected exposures and Rajpurkar et al. did not report testing the CheXNet algorithm on clinically rejected x-rays. Therefore, while the CheXNet algorithm has been shown to successfully classify quality x-rays, it is unknown whether the system can identify pneumonia in classically rejected x-rays, which motivates this study.

2. Data Set

Ideally, data for this project would come from rejected chest x-ray images that ultimately were diagnosed through image retake. Therefore, the rejected images could be be annotated (diseased or healthy) based on the retakes and the CNN would learn on real-world data.

However, open source data-sets of annotated, rejected x-ray images do not exist (or at least could not be found by the authors). Therefore, we have resorted to taking quality x-ray images and applying clinically relevant augmentations to resemble classically rejected x-rays. These augmentations are described in Section 3.1, shown in Figure (1), and represent the causes of the vast majority of image retakes

(Lin et al., 2016).

The original, unaugmented chest x-ray images came from a data set of one- to five-year old pediatric patients from the Guangzhou Women and Children's Medical Center. There are 5,863 images and two categories: pneumonia and normal. Three medical experts were responsible for classifying the images. Low-quality and unclassifiable images were thrown out. The data was initially acquired and used in a 2018 study on identifying medical diagnoses with image-based deep learning (Kermany et al., 2018).

3. Method and Model Development

Like the Stanford Machine Learning Group, our study uses the DenseNet-121 CNN to learn and predict based on the medical imaging data. By reusing a method that proved effective for a similar problem, this study attempts to isolate the effect of clinically relevant data augmentations on state-of-the-art learning algorithms. Furthermore, the fact that DenseNet-121 has already been trained on an extensive set of image data (ImageNet) makes it an easier "plug and go" option for the purposes of a paper of this scale.

DenseNets are distinguished by the use of feature-maps from one layer as inputs in each subsequent layer; consequently, each layer has a set of inputs from each of the preceding layers. Relative to other deep learning techniques, DenseNet makes use of narrow layers (roughly 12 filters per layer).

The model was trained on the non-augmented medical images over 13 epochs with a batch size of 64. This training set of 5216 images was used with a 10% validation split. The model was stopped at 13 epochs to prevent further over fitting because training accuracy and training loss was stalled, as was validation accuracy. The model was then tested on both a set of non-augmented medical images and sets of augmented medical images derived from the same non-augmented test set.

Although the model is not trained as thoroughly as in CheXNet, the non-augmented test set performance still provides a basis from which to evaluate how different augmentations affect predictive accuracy. Should an augmentation not lead to a notable decline in performance, there may be some indication that machine learning could be used to circumvent barriers to human classification.

3.1. Data Augmentation

The original chest images were all of different sizes, varying greatly in either dimension. All x-ray images were then re-scaled to 400x400 pixels using image anti-aliasing methods from the *Image* library. This both established the structural consistency required by the CNN and reduced the data to allow for timely training on standard laptops. Then, the *Imgaug* library was used to perform the following augmentations.

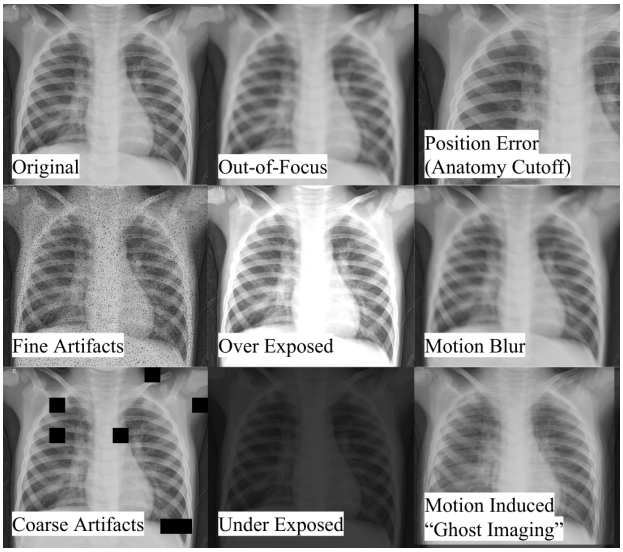

Figure 1. Example of data augmentations performed to represent rejected chest x-rays

3.1.1. Out-of-Focus

The out-of-focus augmentation represents x-rays where the subject of interest was not in the optical focal plane of the machine, leading to normal blurring. It is represented in the model by applying a Gaussian blur to the images. The level of augmentation can be controlled through the mean and standard deviation of the Gaussian distribution.

3.1.2. Position Error

Position errors occur when either the patient is posed incorrectly or the machine is aimed incorrectly. Importantly, these errors lead to sections of the anatomy to fall out of the field of view. To represent position error, images are translated in the x-y plane and then uniformly scaled lager. The axes independent shifting and global factor of scaling are controlable for the augmentation.

3.1.3. Artifacts

Fine artifacts, such as dust, can contaminate the x-ray film. Further, coarse artifacts, such as objects can also occlude imaged areas. In the model, artifacts are represented through random pixel dropout. This can be controlled by both the percentage of pixels to drop out (how much occlusion) and how dense or sparse these dropouts should occur (differentiating the fine and coarse variants).

3.1.4. EXPOSURE

This augmentation represents incorrect exposure settings on the x-ray machine, leading to loss of data by either whiting or blacking out regions of the image. Changes in image exposure were obtained by multiplying pixel values by a constant. Constants greater than 1 overexposed the image while constants less than 1 underexposed the image. [1]

3.1.5. MOTION

The final augmentation represents patient motion during imaging. This augmentation comes in two types: motion blur and ghost imaging. Motion blur represents directional blurring caused by continuous, unidirectional movement during the imaging. Ghost imaging represents step-wise motion that creates a displaced double image effect. Motion blur is represented by blurring pixels of a certain kernel size in a given direction. Ghost imaging is achieved by overlaying a shifted version of the original image and blending with a given alpha blending value. Finally, the kernel size, shift percentage, and alpha blending values are all controllable in the model.

3.2. Further Image Augmentation Work

Further investigations with this data-set could take advantage of known anatomy. The information for diagnosis of pneumonia resides in the lungs, which are at the center of chest x-rays. Therefore, masks could be applied to the images that would effectively make them circles. The preprocessing would remove erroneous information from both natural influences (other parts of the body) and artificial influences (text, labels, or scale bars that often occur in the image). This change would make the CNN more robust as the model could not learn parameters that were not directly caused by pneumonia.

4. Results

In this section we detail the effects of each of the augmentations on the evaluation of our model. This was done by first using a test set of 624 clean x-ray chest images of the same form as the training set to establish a baseline accuracy. The model was able to classify this set with an overall accuracy of 78.7%, with this mostly comprising of a high TPR of 97.9%. Then, the same 624 images were augmented using the methods described in Section 3.1 the run on the model. Note, the augmentations in the test data use the same model parameters used to produce Figure (1). Further each image received a single augmentation.

[1]While the multiplicative exposure method worked well to represent overexposed images (as multiplying by a constant can set a pixel value to its maximum), underexposure would have been better represented by subtraction.

From Table 1, the respective accuracy, true positive rate (TPR), and true negative rate (TNR) of the original dataset and each augmentation can be seen.

Table 1. Predictive accuracy and true positive and negative rates for each augmentation type.

	Accuracy (%)	TPR (%)	TNR (%)
Original	78.7	97.9	46.6
Coarse Artifacts	77.6	97.3	41.0
Fine Artifacts	51.9	28.2	91.5
Out-of-Focus	67.9	98.7	16.7
Ghost Image	73.1	97.4	32.5
Motion Blur	67.8	97.9	17.5
Overexposure	72.3	59.7	93.2
Underexposure	62.5	100	0.00
Position	63.6	94.1	12.8

5. Discussion

The true positive and negative rates for the original test set are characteristic of how the model tended to perform regardless of which augmentation was applied. The true positive rate (TPR) of 97.7% and the true negative rate (TNR) of 46.6% on clean images imply that the model is quick to classify pneumonia. Part of this tendency may follow from the fact that the majority (75%) of the training data is positive for pneumonia. Of the test set, however, only 63% of the images are classified as pneumonia. While further developments would balance the data sets, it is still clear the model is not just blindly classifying each image as positive.

5.1. Artifact Distortions

The coarse artifact augmentation did not have an enormous impact on model performance. The TPR remained similar (97.3% from 97.9%) while the TNR dropped a little more steeply (41.0% from 46.6%). Much of the image is usually unaffected by coarse artifacts and it would take bad luck for the blocks to cover all the parts relevant to a pneumonia diagnosis. However, it is important to recognize that general performance does not indicate perfect performance for an individual patient. While a radiologist may be able to determine which specific artifacts may affect a diagnosis, our model cannot.

The fine artifacts augmentation is a different story. It is one of two augmentations where the TPR and TNR seem to switch roles (TPR at 28.2% and TNR at 91.5%). It seems that there is something about the evenly distributed missing data that makes the model unlikely to classify an image as positive for pneumonia. The consequence here is that the overall accuracy drops to a very low 51.9%, indicating almost incomplete ineffectiveness of the model. Importantly, these augmentations do not look very severe to a human

(see Figure (1)), so radiologist may have overconfidence applying these images to a CNN similar to ours.

5.2. Focus and Blur Effects

The out-of-focus augmentation significantly increases the false positive rate. Radiologists detect pneumonia by looking for white spots in the lungs caused by the infection, so it is possible the blurring effect brings the white of the bones in to the lung cavity, creating light areas that appear as pneumonia to the classier. The ghost image augmentation seems to have a similar, but reduced, effect. While it creates a similar blur, it is possible that it tends to wash out fewer important details than the out-of-focus augmentation. Finally, the motion blur effect is in-line with the out-of-focus augmentation in terms of impact. This implies that it washes out a similar amount of data.

Overall, the performance on blurred images–especially ghost images–outperformed the fine artifact detection rate. This is surprising as the blurring appears worse than fine artifacts to human intuition, indicating a possible application for CNN use.

5.3. Changes in Exposure

The exposure tests provide further insight to the models classification methods. Namely, decreasing pixel brightness (underexposure) caused the CNN to classify every image as diseased while increasing pixel values (exposure) increased the model's affinity to predict healthy images. Overall, it appears the model associate dark values with pneumonia and light values with health, which is opposite of the expected result (as pneumonia creates light patches). However, another way to view the model is that the underexposure decreases the gradient between pixels while the overexposure increases the gradient (assuming values are not whited out). This gradient difference may also indicate the way the model learns, showing smaller gradients are more likely to be diseased.

5.4. Position

The position augmentation significantly increases the false positive rate, with the TNR at 12.8%. The results of the TNR are reasonable, as cropping of healthy images simply removes overall information of the images. In unhealthy images, cropping can result in concentration of the image on the affected area, thus magnifying the image to the point of interest. When the cropping works unfavorably, and crops out part of the affected area, there is a loss in accuracy.

5.5. Error and Further Work

As described, the model was trained only only clean x-ray images as we were interested on its performance to novel augmentations. However, if the model was trained with augmented images as well (even different form the tested augmentations), the model may have been more robust to the clinical augmentations. For example, if the model had also been trained on "inverse" images (switching the light and dark regions), it may have classified more strongly on contours than on colors and been more robust to the exposure experiments.

Additionally, as mentioned, the unbalanced data set provided incentive for the model to detect pneumonia. While this ultimately created a greater false positive than false negative rate–which is often desired in the medical community–a more robust model would be trained on balanced data.

Further work would implement the pre-processing described above and would also furhter explore the parameter space of the augmentations to learn more about the limits of detects. Advances could even take lessons learned from adversarial attack research and attempt to pre-process test images to "de-augment" them. Fianlly, retraining of the model would be conducted on Google Colab to utilize GPUs to allow for more epochs per time than we achieved on our CPUs.

6. Conclusion

From this work, we conclude that naive implementations of CNNs generally fail to detect pneumonia at high accuracy rates in images distorted in ways that would require retake. Specifically, the inclusion of fine artifacts, changes in exposure, and position errors have the greatest effect on classification accuracy, while blurring effects have lesser impacts on accuracy. These results are interesting as the CNN performed worse in areas where humans see less distortion (exposure, fine artifacts, positioning) and better in the areas where humans see greater distortion (blurring effects). Not only does this indicate properly trained CNNs may have a role to play in classifying blurred images, it also warns against naive application of CNNs to medical problems that appear simple to humans that may be more complex to machines. Overall, the study reinforces training models on augmented data and provides further motivation to build CNNs that can detect pneumonia on degraded images as it may serve as a solution to limit patient radiation exposure.

Acknowledgments

This project uses data from Kaggle.com originating from a study title *Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning* by (Kermany et al., 2018). The use of DenseNet121 to tackle this issue is inspired by the similar usage of the same model by (Huang et al., 2017) in 2017.

References

Berkhout, WE. The alara-principle. backgrounds and enforcement in dental practices. *Nederlands tijdschrift voor tandheelkunde*, 122(5):263–270, 2015.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Huang, Gao, Liu, Zhuang, Van Der Maaten, Laurens, and Weinberger, Kilian Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Kermany, Daniel S, Goldbaum, Michael, Cai, Wenjia, Valentim, Carolina CS, Liang, Huiying, Baxter, Sally L, McKeown, Alex, Yang, Ge, Wu, Xiaokang, Yan, Fangbing, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

Lin, Chih-Sheng, Chan, Po-Chou, Huang, Kuang-Hua, Lu, Chun-Feng, Chen, Yung-Fu, and Lin Chen, Yun-O. Guidelines for reducing image retakes of general digital radiography. *Advances in Mechanical Engineering*, 8 (4):1687814016644127, 2016.

Rajpurkar, Pranav, Irvin, Jeremy, Zhu, Kaylie, Yang, Brandon, Mehta, Hershel, Duan, Tony, Ding, Daisy, Bagul, Aarti, Langlotz, Curtis, Shpanskaya, Katie, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Wootton, Dan and Feldman, Charles. The diagnosis of pneumonia requires a chest radiograph (x-ray)—yes, no or sometimes? *Pneumonia*, 5(1):1, 2014.