InceptionNet CNN for X-ray Pneumonia Classification

Ang Jie Liang (A0149293W), Chua Kang Wei (A0173557W), Frederick Liew Yong Lun (A0168902Y), Seow Zhi Xian Gabriel (A0185064A), Richard Wong Ho Chuan (A0181320W), Wong Tau Yew, Mark (A0190753B)

CS3244 Group 3

Abstract

This paper investigates the use of convoluted neural networks (CNN) to predict and classify the presence of pneumonia from X-ray images. We propose the use of InceptionNet architecture which is able to produce a heat map of the X-ray which can highlight regions that are indicative of pneumonia so that radiologists can make well-informed decisions instead of just relying on a final classification.

Introduction

Pneumonia is an inflammatory condition of the lung which can be life threatening if not diagnosed in time. According to the World Health Organisation (2019), pneumonia makes up for 15% of all deaths of children below 5 years old. Chest radiology is critical in the diagnosis of pneumonia as it is able to detect the presence of pneumonia and its severity (Wootton 2014). However, X-ray image analysis is a tedious task for radiology experts. Machine learning (ML) solutions can alleviate these problems, by reducing radiologies' workload as well as providing further and more accurate insights from the X-ray images. In light of the COVID-19 situation, where WHO (2020) has stated that the most common diagnosis for patients with severe COVID-19 is severe pneumonia, it is even more critical to reduce the strain on healthcare personnel and resources through ML.

As such, our application aims to cull medical workers' workload by providing a baseline interpretation for radiologists to work off on, potentially reducing misdiagnosis. Our ultimate aim is to improve the quality of life for healthcare workers as our application is a force multiplier, expanding the accuracy and speed of radiology pneumonia diagnosis.

Related Work

The research area of medical image classification is increasingly popular as researchers propose different ML algorithms to detect and classify thoracic diseases. Antin et. al. (2017) proposed a logistic regression model that provides a binary pneumonia classification based on chest X-ray images. Rajpurkar et. al. (2017) developed CheXNet based on a 121-layer DenseNet architecture which classifies 14 pathology classes including pneumonia. Jaiswal et al. (2019) built Mask-RCNN, a deep neural network that utilizes pixel-wise segmentation to identify pneumonia. Sirazitdinov et. al. (2019) proposed an ensemble of two CNN, RetinaNet and Mask R-CNN for pneumonia detection and localisation. Many of the existing researches are based on large datasets and apply different CNN architectures to output a binary classification of pneumonia. In this paper, we propose the use of Inception CNN network to not just accurately classify pneumonia, but also to provide additional output data such as heatmap localisation, variance and confidence intervals to assist medical experts in deducing from our model results.

Dataset and Data Preprocessing

Our datasets consists of chest x-ray images that only contain the erect anteroposterior chest view.

Training and Validation sets are derived from a common pool taken from "COVID-19 Radiography Database" by Tawsifur Rahman (2020) from Kaggle.

It consists of normal and pneumonia images in an almost 1:1 proportion(1341 : 1345). A separate independent Test set is created from a set of COVID 19 images found in a separate dataset (Cohen 2020), and mixed with normal x-ray images found in another dataset "Chest X-Ray Images (Pneumonia)" by Paul Mooney (2019) in Kaggle to test the model on.

In terms of data preprocessing, we tried to segment out the boundaries of the lungs. This will increase the validity of predictions since these will be based on the lungs and not on other irrelevant features/organs such as the heart or the neck. The procedure used comprises four stages: image enhancement, thresholding, noise removal and contour filtering.

Methods used to improve contrast are Balance Contrast Enhancement Technique (BCET), Contrast Limited Adaptive Histogram Equalization (CLAHE) and Local Contrast Enhancement (LCE). For thresholding, Otsu's Thresholding is used to obtain a binary threshold value robust to dynamic lighting conditions. The morphological opening operation and median filter are found suitable for noise removal. With enhancement based on a subcombination of the above mentioned techniques, the image is then divided into half, shapes at the boundary of the image removed, and shapes with the largest area of each half are taken to represent the lungs. Finally, a convex hull may be drawn and taken as boundaries of the lungs. Upon inspection, the lung boundaries of a number of images after processing are either truncated or non-existent either due to presence of horizontal rib bones leading to lung regions being splitted or poor separation between the image boundary and the ribcages leading to the entire areas being filtered out. Therefore, we decided not to use lung segmentation for our proposed method due to its inconsistency.

Instead, data augmentation in the form of random shifting, scaling, zooming and horizontal flipping is performed on images before being inputted into our model for training. This allows for us to train our model over various transformations of the same image, helping training to converge on the prominent features at a faster rate.

Initial Work

Other than convolutional neural networks, some examples of ML models based on images are support vector machines (SVM), K-nearest neighbours (KNN) and logistic regression (LR). Selected features are required to be extracted from the images before fitting to these models. In this case, our chosen features are properties of the Gray Level Co-occurrence Matrix (GLCM) and Haralick features. These are texture features which are scale-invariant and are commonly associated with classification of various diseases in the medical field i.e MRI images of liver, heart, prostate and brain tumour. Here is a summary of the performances of these models.

Table 1:

Model	Training accuracy	Validation accuracy
SVM (1, 2, 3, 7, 10, 13, 17, 20, 23)	0.955	0.961
KNN (5 neighbours) (1, 3, 17, 25)	0.952	0.953
LR (1, 3, 12, 13, 14, 15, 17, 18, 20, 25)	0.922	0.929

Training is not done on all combinations of the 26 texture features. Presented for each model (in brackets) is a combination of features with the "best" validation accuracy known to us.

Variables 0-12 are GLCM properties while variables 13-26 are Haralick features. Notably, variable 1, 3 and 17 are the dissimilarity of GLCM, energy of GLCM and the 5th Haralick feature (average inverse difference moment) respectively. 20% of the dataset is used for validation. However, we will be focusing on CNN on the sections below since it is able to demarcate the critical regions of the images used in prediction while these regions are hard to tell from the values obtained as texture features.

Methodology

Through studying popular CNNs architectures such as VGGnet and InceptionNet, we observed that many used deep neural networks, with a large number of convolutional layers to increase non-linearity and improve generalisation power. As a result, many of these architectures are computationally demanding and have undergone extensive tuning of hyper-parameters. These networks also implemented many techniques (e.g. aggressive dropout, batch normalization and auxiliary classifiers) to deal with overfitting and the vanishing gradient problem, problems prevalent in deeper neural networks.

When designing the model, our aim was to achieve a simpler and more 'user-friendly' version of the InceptionV4 architecture that required lesser training time. Hence, we focussed on designing a robust model that retained key features from InceptionV4, while excluding features that were unnecessary in a shorter network.

Our model primarily consists of several alternating inception modules and reduction blocks, with a final Global Average Pooling (GAP) layer, followed by a fully connected layer. The final output of the model is given by a sigmoid activation unit.

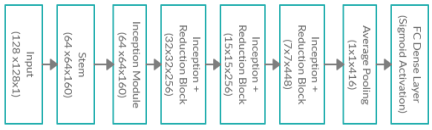


Figure 1 Model Architecture

The model uses batch normalisation to perform regularization in the convolutional layers and fully connected layers, as an alternative to aggressive dropout and auxiliary classifiers used in InceptionNet. Batch normalisation was chosen as it has been shown to significantly reduce the number of training epochs required (Brownlee 2019) while outperforming dropout in convolutional neural networks (Jansma 2019). During training, we also observed that using dropout in the fully connected layer caused large fluctuations in validation loss and validation accuracy.

The model implements tweaks to the original inception modules and reduction blocks to significantly reduce the filter dimensionality throughout the network. This follows from the principles suggested in the original Google InceptionNet paper (Szegedy 2014), for reducing computation time. The model uses the simplified reduction blocks proposed in InceptionV4 and the inception module proposed in InceptionV1, while excluding the max pooling layer.

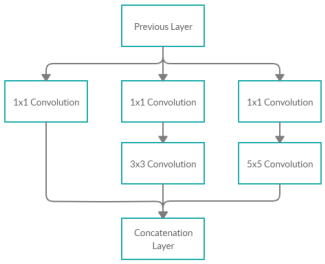


Figure 2 Inception module

In figure 2, we have the model's implementation of the inception module. The module performs convolutions of varying kernel sizes (i.e. 1x1, 3x3, 5x5) on the same input and outputs a concatenation of the layers produced from these separate convolutions. By adjusting the proportion of filters for convolutions of a specific size, we can adjust the 'weightage' of certain features. For instance, we increase the proportion of filters allocated for the 5x5 and 3x3 convolutions to increase the weightage of larger features over smaller localized features. During training, this gives us greater control over the type of features that the model would focus on, instead of the typical approach of using more ambiguous CNN training heuristics.

The model uses a weighted variant of the binary cross entropy loss function, where loss calculated from false negatives is scaled higher than loss from false positives. This weighted variant was used as we observed that the model had significantly more false negatives than false positives especially in later epochs of training. This could be due to it being more difficult for the model to identify 'normality' in x-rays as compared to identifying indicators of pneumonia.

As the model uses GAP as the final pooling layer, it performs object localisation and allows us to generate a Class Activation Map (CAM) from the original x-ray (Cook 2017). CAM is a heatmap that highlights areas of the x-ray that contribute to the

model's prediction of pneumonia or 'normal'. For instance, when the model's prediction is pneumonia, the corresponding heatmap reflects areas on the x-ray which the model identified as indicators of pneumonia, and similarly for 'normal' classifications. During training, we took a random sample of images to determine if image noise or artifacts in the image (e.g. instrument readings) were highlighted and contributed to the model's prediction. This allowed us to gauge our models' resilience to noise and additional artifacts and to implement corresponding measures to reduce overfitting. The figures below also show the models ability to correctly identify regions within the chest boundary.

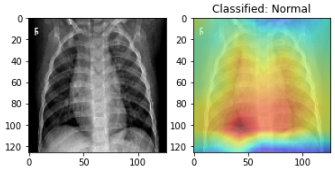


Figure 3 Normal x-ray (CAM)

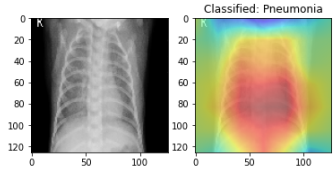
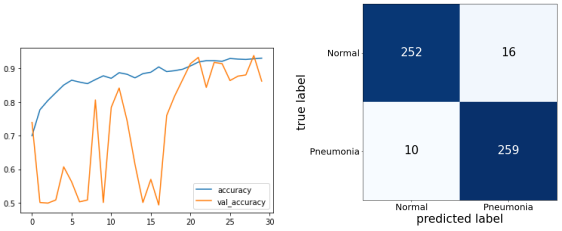


Figure 4 Pneumonia x-ray (CAM)

Results

a. Model Performance

The model performed well, achieving a peak validation accuracy of 94.4% while reflecting a stable validation accuracy after roughly 18 epochs. Notably, weighted binary cross entropy contributed an approximate 2% increase in validation accuracy across different models.

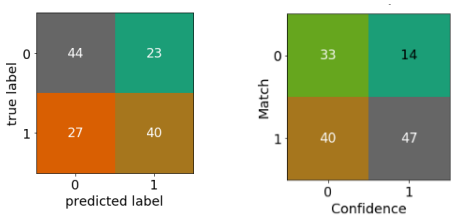


Left: Figure 5 Accuracy and Validation Accuracy

Right: Figure 6 Confusion Matrix on Test Data

b. Testing with COVID 19 dataset

In light of the COVID 19 epidemic as of the time of the submission of this paper (4 May 2020), we have decided to test our model on a Test set consisting of 67 SARS-related coronavirus pneumonia images and 67 normal images.



Left: Figure 7 Confusion Matrix of Actual vs Predicted

Right: Figure 8 Confusion Matrix of match/mismatched labels between actual and predicted labels, vs Confidence levels

The predicted values compared to the actual test values (refer to Figure 7) only yield an accuracy of 0.6567, which is a degradation from the validation set. However, by estimating a 95% confidence interval for each predicted value (explained in Application), we are able to differentiate predictions of high confidence (95% CI does not include 0.5) and low confidence (95% CI includes 0.5). For predictions with low confidence, the model is more likely to misclassify since values are closer to the 0.5 threshold and also have larger variances. This can be seen in Figure 8 where Confidence=0 corresponds with an almost even mismatch/match ratio, suggesting that the model is unsure of these values around the 0.5 threshold. On the other hand, predictions with higher confidence are more likely to be classified correctly with a much lower mismatch/match ratio of 14:47.

Application

We propose an application which applies our Inception CNN network to aid radiologists in detecting pneumonia. The application is a script that takes in a directory of x-ray image data, and outputs a CSV file (refer to Figure 7.1) containing the following information: name of the files, the class of the image (normal or pneumonia), the expected probability output of the neural network, the variance of the associated probability value, 95% confidence interval, and a heatmap highlighting areas of the x-ray which contributed to the model's prediction.

It should be noted that the network outputs only a single float value as a result of the single sigmoid unit found in the last layer. Though we can use a >0.5 condition to determine if this value tells us if its a normal (0) class or pneumonia (1) class, the sigmoid output cannot be used to meaningfully extract a measure of uncertainty of the prediction. Fortunately, a method was proposed by Yarin Gal (2015) to address the issue of uncertainty in neural network predictions using a Gaussian Process approximation. This involves the use of a dropout layer in the model after the last inner-product layer. By applying the dropout in the prediction process, it simulates a sample from the network, and by taking the mean and variance of the samples of a prediction, we are able to achieve the expected model output given an input image (predictive mean), and the confidence of the model in the expected prediction value (predictive variance).

The output data sheet of the application from a given set of image inputs would therefore have a measure of the confidence interval of the prediction. Predictions with intervals not containing 0.5 would therefore be classified as a high confidence prediction and vice versa. Therefore Healthcare workers are able to interpret the model output and differentiate between high confidence predictions and low confidence predictions, allowing radiologists to prioritize focus on the lower confidence images.

	filenames	True Labels	Predicted Labels	Predictive Mean	Predictive Variance	Matching Labels?	95% interval_lower	95% interval_upper	is 0.5 in CI?	High Confidence?
0	NORMAL\IM-0001-0001.jpeg	False	False	0.447547	4.933756e-02	True	0.012191	0.882903	True	False
1	NORMAL\IM-0003-0001.jpeg	False	True	0.557017	7.338075e-02	False	0.026075	1.087959	True	False
2	NORMAL\IM-0005-0001.jpeg	False	False	0.363000	2.771908e-02	True	0.036678	0.689321	True	False
3	NORMAL\IM-0006-0001.jpeg	False	True	0.752425	4.212939e-02	False	0.350126	1.154723	True	False
4	NORMAL\IM-0007-0001.jpeg	False	False	0.452339	5.583793e-02	True	-0.010810	0.915488	True	False
...	...	...	...	...	...	...	...	...	...	...
129	PNEUMONIA\covid-19-pneumonia-rapidly-progressi...	True	True	0.998324	4.339319e-06	True	0.994241	1.002407	False	True
130	PNEUMONIA\covid-19-rapidly-progressive-acute-r...	True	True	1.000000	6.394885e-14	True	0.999999	1.000000	False	True
131	PNEUMONIA\covid-19-rapidly-progressive-acute-r...	True	True	0.850000	3.872719e-02	True	0.464287	1.235713	True	False
132	PNEUMONIA\covid-19-rapidly-progressive-acute-r...	True	False	0.239446	1.566532e-01	False	-0.536311	1.015203	True	False
133	PNEUMONIA\covid-19-rapidly-progressive-acute-r...	True	True	1.000000	0.000000e+00	True	1.000000	1.000000	False	True

Figure 9 Demonstrating Output of Application

In the context of radiology, an accurate diagnosis is critical to ensure that the individual receives the appropriate treatment. However, studies have shown that even robust models are prone to misclassifications when images include additional artifacts (Eykholt 2018). Hence, a large concern is whether the model is still able to give an accurate prediction on images containing artifacts and anomalies.

In radiology, this is often the case, as images can contain additional artifacts, ranging from instrument readings, “L” and “R” markers to indicate the orientation of the image, to feeding tubes and pacemaker implants. By providing radiologists with a heatmap, they can assess the accuracy of a prediction by observing whether an artifact had interfered with the model’s prediction.

Ethical Issues

ML algorithms should utilise anonymous data that respect the privacy of patients since we are using personal health data. Furthermore, since these algorithms are fed with large volumes of health data, a large scale data leak will infringe on the privacy rights of patients as they can be easily misused.

Moreover, we should use representative data for training and validation, especially when different demographics of the population may exhibit different medical data, causing the ML algorithm to develop biases. For example, it is well documented that ML can lead to conclusions that disfavours the poor or certain races, thus leading to social or racial discrimination. Hence, the ML algorithm has to be trained on a suitable demographic set. This also means that the algorithm developers need to relay the workings of the algorithm to the doctors for them to make meaningful decisions which means the algorithm cannot be a black-box. As such, the heat-map of our model allows radiologists to make informed decisions, and observe if there are any foreign objects, such as feeding tubes, that may cause a misclassification.

Furthermore, the doctors also have to communicate this information to the patient so that there is clear transparency and accountability for both parties to make decisions. The patient should be prepared for the possibility of false diagnosis conclusions from the ML algorithm. The doctors themselves also cannot fully rely on the ML algorithm as there may still be biases or unique cases that are not accounted for by the algorithm. This is especially important since the welfare and livelihood of the patients are at stake, and they should have the final say to trust the ML algorithm or to go for further tests, which may result in higher expenses.

Conclusion

We managed to attain competitive test scores and apply it to novel data with fairly good predictions (for high confidence predictions). A key issue with the model is that each training session takes 30 minutes with the use of a high-end GPU. This is in spite of using a model that is already much leaner than competitive models. We identified several key areas where we could improve the model performance: improving on image pre-processing techniques (e.g. cropping into a lung via boundary identification), inclusion of other models identified in the Initial Work via the use of an ensemble network architecture. In conclusion, we have demonstrated the use of InceptionNet architecture in CNN in order to better classify and diagnose pneumonia via x-ray images to assist in healthcare efforts.

Personal Contributions and Learning Reflection

Richard explored how to include uncertainty in model predictions, and generalise the model to other data sets, both needed to adapt the model well to the given application function. Richard learned the importance of uncertainty of predictions in neural network models.

Mark explored the segmentation of lungs and other ML algorithms. These constitute preprocessing and initial work done before a choice made to focus on CNN. Mark learnt about the difficulties in selecting what features should be considered and how CNN potentially is able to circumvent this problem.

Frederick bridged technical specifications and real life expectations by doing research into how analysis of chest x-ray images are used by doctors to determine what kind of direction the project should head in to result in realistic benefits. Frederick learnt more about how real life requirements can be used to adjust parameters to acquire a more useful model.

Kang Wei explored regularisation techniques such as elastic net, L1, L2 regularisation to reduce overfitting and examined ethical issues in ML. Kang Wei learnt about further CNN architectures and techniques.

Jie Liang explored the use of CNN architectures such as VGG-16 and ResNet50, and learnt about different novel CNN architectures from existing research papers and publications.

Gabriel constructed the model's InceptionNet architecture and implemented the techniques used to increase regularization, decrease training time and reduce false negatives. Gabriel learnt about the difficulties and tradeoffs when implementing new techniques to achieve good performance.

References

Alexis Cook. 2017. Global Average Pooling Layers for Object Localization.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. 2015. Going Deeper with Convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA: 1-9.

Harrison Jansma. 2019. Don't Use Dropout in Convolutional Networks.

Jaiswal, A.K.; Tiwari, P.; Kumar, S.; Gupta, D.; Khanna, A.; Rodrigues, J.J.P.C. 2019. Identifying pneumonia in chest X-rays: A deep learning approach. Meas. J. *Int. Meas. Confed*. 145, 511–518

Jason Brownlee. 2019. A Gentle Introduction to Batch Normalization for Deep Neural Networks.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song. 2018. Robust Physical-World Attacks on Deep Learning Models. arXiv:1707.08945 [cs.CR].

Paul Mooney. Chest X-ray Images (Pneumonia). 2019.

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

Sirazitdinov, I.; Kholiavchenko, M.; Mustafaev, T.; Yixuan, Y.; Kuleev, R.; Ibragimov, B. 2019. Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database. Comput. Electr. Eng. 78, 388–399

Tawsifur Rahman. 2019. COVID-19 Radiography Database.

Joseph Paul Cohen; Paul Morrison and Lan Dao. 2020. COVID-19 image data collection. arXiv:2003.11597[eecs.IV]. Cornell University Library.

Wootton, D., Feldman, C. The diagnosis of pneumonia requires a chest radiograph (x-ray)—yes, no or sometimes?. *Pneumonia* 5, 1–7 (2014).

World Health Organisation. 2019. Pneumonia.

World Health Organisation. 2020. Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected.

Yarin Gaal. 2015. What My Deep Model Doesn't Know.