
Pneumonia Detection on Chest X-Rays

Gaurav Singh
UID: 305353434
December 6, 2022

Abstract

Pneumonia is a life threatening disease especially when pandemic like COVID-19 can induce it and become a cause for large scale deaths. During times like these there is a need for fast and accurate diagnosis using Chest X-rays to save much more lives. In this paper I have explored different Machine learning and Deep learning methods to classify a patient as Pneumonia positive or negative based on provided Chest X-ray. A comparison of these approaches based on different evaluation metrics and on their run time is shown. Though I have started with simple approaches for comparison, focus is given to State of the art approaches which are Deep Convolution Networks and Vision Transformers.

**The code for this project can be found here: [1]

1 Introduction

Pneumonia is an acute pulmonary infection which causes inflammation of air sacs called Alveoli. It is most common in underdeveloped or developing nations which have below average air quality and sanitation standards. Pneumonia can be of these three forms viral, bacterial or fungal which makes it airborne and contagious. It mostly effects children under the age of five or elderly people who have weak immune system [2]. Symptoms include coughing, fever, difficulty in breathing, loss of appetite etc. and is life threatening. COVID-19 induced Pneumonia has been a major reason for pandemic deaths in the world [3]. Pneumonia is a rapid attacking infection whose quick diagnosis is important for saving lives. Detection of Pneumonia normally requires a highly skilled doctor to look at the Chest X-rays (CXR) and CT scans and requires medical history to arrive at a conclusion. Normally a foggy CXR with white spots are a few indicators of Pneumonia Fig. 1. This manual process might be expensive and critically time consuming. The recent advancements in Machine learning and Computer vision might prove really helpful in quick and accurate detection of Pneumonia and can potentially save a lot more lives especially during pandemics like COVID-19 which faced shortage of human expertise.

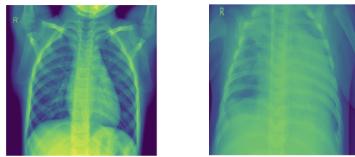


Figure 1: Normal vs Pneumonia infected CXR

2 Data

The data used for this project is based on data originally provided by Kermany et al [4] which is available as open source on Kaggle [5]. There are a total of 5,856 CXR(anterior-posterior) images from pediatric patients of one to five years old. The raw data is organized into three - train, test and validation - folders. Each folder is further having images into two sub-folders namely Pneumonia and Normal. There are 5,216 train, 624 test and 16 validation images originally which I redistributed using Stratified sampling so as to get 600 images for both train and test data which is around 10% of total data. Class distribution of data is shown in Fig. 2 which tells that there is a heavy class imbalance

with $\sim 75\%$ images belonging to positive patients. The images are of varying dimensions therefore all images are re-scaled to 256 X 256 while read. It is noticed that there is channel inconsistency among images with 251 images being RGB whereas 4,381 images being gray-scale. All images therefore are read as gray-scale to maintain consistency.

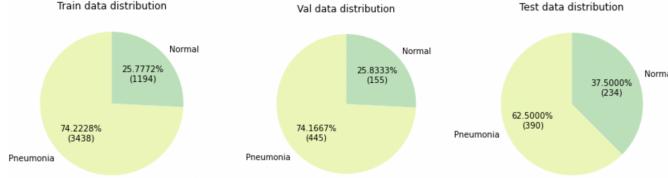


Figure 2: Class Distribution

3 Unsupervised Machine Learning

For the start, I investigated the two classes and is there any significant difference in the data distribution that can be segregated into two groups. For this a relevant work is already done for Brain tumor segmentation using Principal components based clustering [6]. I took inspiration from this paper and used SVD decomposition instead of EM based PCA to find principal components which retain 98% variance. Each train image is unrolled into a 65,536 dimension row vector and I kept 1,648 principal components out of it for faster computation. K-means clustering with 2 clusters, gave two centroids which are then used to classify the test images based on Euclidean distance from cluster centroids. The train imageset is visualised in 2 dimensional space using t-SNE components instead of PCA since it's a non-linear transformation unlike PCA which preserve the Local structure of data by minimizing the Kullback–Leibler divergence (KL divergence). The visualisation Fig. 3 of the train data shows that there is a significant difference between the two class images and thus can be segregated using a good decision boundary.

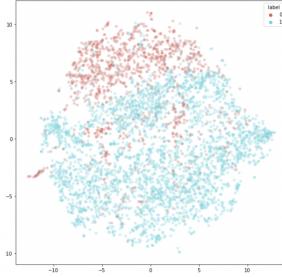


Figure 3: t-SNE plot of train images

4 Supervised Machine Learning

4.1 Logistic Regression

Since Pneumonia detection is a binary classification task Logistic regression is a good baseline to start with. It is just like linear regression with values mapped to probabilities using Sigmoid function and it uses a log loss or binary cross entropy loss function for training. For this approach, Principal components of images are used instead of full length data. Regularization value of 0.001 is used to avoid over-fitting and class weights are balanced by a factor of inverse of no. of per class samples to deal with imbalanced dataset.

4.2 Support Vector Classifier

Logistic regression decision boundary is linear and it's shown in Fig. 3 that a non-linear decision will be a good choice. I used Soft margin Support Vector Classifier ($C = 1$) with Radial Basis Function (RBF) as kernel to get non-linear decision boundary. Principal components are used instead of full data and the weights are balanced here as well.

5 Convolution Networks

Convolution networks provided the major shift in the computer vision tasks by using the concept of filters which go through the full image and generate feature maps using element wise matrix multiplications. They have less number of parameters to work with given filters have less trainable parameters and full connections are limited to last few layers.

5.1 LeNet architecture

This is the first CNN architecture which uses 4 [CONV-POOL] layer and 3 Fully connected(FC) layers as shown in Fig. 4(a). The network is modified a bit with Batchnorm and ReLU as activation function. Batchnorm helps to manage Internal Covariate Shift which enables us to use higher learning rate and has regularization effect.

5.2 DenseNet-121 architecture

DenseNet-121 is a 121 layered CNN which was used by Rajpurkar et al. [7] for pneumonia classification. DenseNet uses dense blocks as shown in Fig. 4(b) to pass all previous layers outputs to current layer to propagate information to great depths and solves the problem of vanishing gradients. To keep the dimensions in control from all these previous output concatenations, 1x1 convolutions in Transition layers are used between each Dense Block.

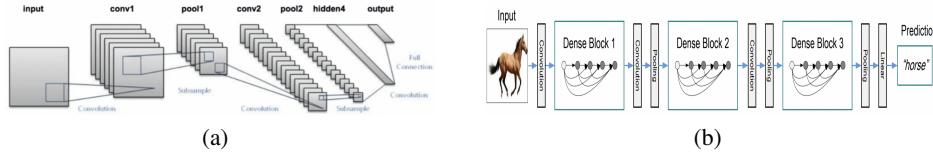


Figure 4: (a) LeNet-5 (b) DenseNet-121

6 Vision Transformers (ViT)

CNNs are gradually being replaced by the State-of-the-art Vision Transformers proposed by Dosovitskiy et al [8] because of their overall computation efficiency, easy GPU parallelisation, ViTs don't require large number of layers as CNN for comparable feature representations and have greater receptive field. They incorporate skip connections in each encoder and decoder and thus are not susceptible to the problem of vanishing gradients. Transformers are used in Natural Language Processing(NLP) tasks but if we break the image into patches which are analogous to words and flatten these patches which are analogous to word embeddings and feed it to a transformer we get a vision transformer.

6.1 Deep Vision Transformer

Deep ViT proposed by Zhou et al [9] as shown in Fig. 5(a) is just like a basic ViT but with increased number of encoder blocks. I have used 12 such encoder blocks with 24 multi attention heads to compute more feature maps. The patch size used is 32.

6.2 Compact Convolution Transformer (CCT)

CCT proposed by Hassani et al [10] as shown in Fig. 5(b) uses compact transformers by using convolutions instead of patching and performing sequence pooling for reducing trainable parameters. This allows for CCT to have high accuracy and a low number of parameters. Also using a mix of both convolution feature maps and simple patches as input reduce the inductive bias and helps transformer to learn better representations. I used 6 convolution layers with kernel size as 3 x 3 to get initial feature maps and 6 multi attention heads and 14 encoder layers.

6.3 Cross Vision Transformer

Chen et al [11] proposes to have two vision transformers processing the image at different scales as shown in Fig. 5(c), cross attending to one every so often. They show improvements on top of the base vision transformer. I have used the upscaled dimensions as 384 x 384 and downsampled dimensions as 192 x 192 with 8 multi attention heads and 3 encoder layers.

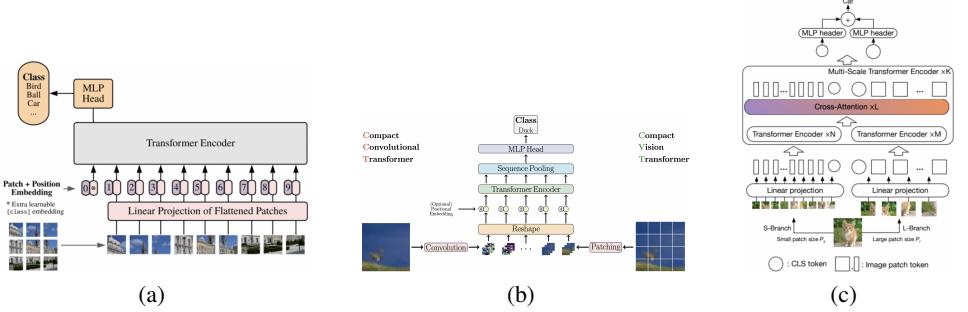


Figure 5: (a) Deep ViT (b) CCT (c) Cross ViT

7 Hyperparameters, Initialisation and Training strategy

Unsupervised methods and supervised machine learning approaches are trained on a 16 core CPU machine whereas the Deep learning approaches i.e. CNNs and ViTs are trained and tested on 6 Nvidia Tesla K80 GPU using the data parallelism approach for fast, efficient and distributed training. For Machine learning approaches i.e. Logistic regression and SVC best hyperparameters (kernel type, regularization coeff.) are found using grid search methods.

For Deep Learning approaches, Xavier initialisation is used for fully connected layers and Kaiming initialisation is used for Convolution layers. For DenseNet-121, the initial weights are taken from model pre-trained on Imagenet dataset. Loss function used during training is cross entropy to which the softmax outputs are fed. Unsupervised, Machine learning and LeNet architecture is trained on single channel images whereas Deep Learning(DL) architectures are trained using 3-channel images formed by stacking the same grayscale image three times.

For DL methods, each image was rescaled to a dimension of 224 X 224 from 256 X 256 and since the training data is less and given DL training is data intensive, each image is horizontally flipped, scaled shifted (by a factor of 0.1) and rotated (by a factor of 5 degrees) with a probability of 0.5 to augment the training data. Data augmentation helps in better learning and helps in avoiding over-fitting. Each image is then normalised by using the mean = 0.48 and standard deviation = 0.22 which are found from the train images. Different learning rate scheduling methods are used with different DL approach. For ViT, 10% of total steps are used as warm-up steps with learning rate of 1e-4 and linearly decayed scheduler is used. For DenseNet-121, cosine decay scheduler is used with learning rate of 1e-4. It is noted that transformer models don't train well i.e. loss doesn't decrease, if warm-up steps are not used. Adam optimizer is used for training.

8 Results discussion and Conclusion

Model	Acc%	Precision	Recall	Train Time(s)	Test Time(s)
PCA + Clustering	58.17	68.48	61.28	279.53	2.21
Logistic Regression	71.16	68.04	94.46	105.65	0.035
SVC	75.00	71.64	93.94	5.99	0.803

Table 1: Results: ML Approaches

Model	Acc%	Precision	Recall	Epochs	Train Time(s)	Test Time(s)
Mod. LeNet	77.40	73.44	1.0	50	150	3
DneseNet-121	83.05	79.80	99.74	50	4000	16
Deep-ViT	80.60	78.01	99.23	20	3060	13
CCT	84.08	81.17	1.0	20	1200	7
Cross-ViT	88.25	84.02	99.42	20	3780	18

Table 2: Results: DL Approaches

From Table 1, we can see that the ML approaches didn't do well in terms of our evaluation metrics i.e Accuracy, precision and recall. For medical diagnosis, high Recall is much more important i.e. to

minimize the False Negatives, also a high Precision is desired, given too many False Positives can lead to wastage of time in critical times like COVID-19. PCA + Clustering gave poorest results and took the most amount of train and test time due to PCA decompositiion. Logistic regression and SVC did well with a high recall but still >20 False negatives as seen from Fig. 6. One thing to note is that SVC can't be used with large no. of Principal components or with large no. of training samples given its runtime is cubic in the max(samples, features). One reason that ML approaches are not able to do well is that by flattening an image we loose spatial structuring which is important for any image identification. Also the decision boundary of LR is linear whereas for SVC is not that complex to segregate the subtle image variations.

From Table 2 we can see that ViTs did well as compared to CNN models in terms of all metrics giving high precision and recall, with less no. of epochs to get the same or even better results. DenseNet performed well but it has a large number of layers and it uses pretrained weights from Imagenet which are not available off the shelf for ViTs yet. A standout metric for DL methods is that, recall is very high which is very much desired as seen from Fig. 7 and Fig. 8. Among CNNs, DenseNet achieved best results by far with 83.05%accuracy and a very high recall of 99.74%. It would be worth to try more complex architectures like ResNet-50 or InceptionV2.

It's quite fascinating how well ViTs performed even without any pre-training and given that Transformer models only start to do well after they are trained for large no. of epochs on large relevant datasets. Cross-ViT achieved the best results 2 with 88.25% accuracy and 99.42% recall. The next step can be to pretrain the ViTs on Imagenet dataset and then to fine tune the model on Pneumonia dataset. From Fig. 7(c) we can see that Deep-ViT has around 127M parameters but the results are not good when compared to Cross-ViT having only 75M parameters. Similar thing can be seen between LeNet(9.7M) and DenseNet(6.9M). Thus, having a large no. of parameters will not guarantee towards better results, the choice of DL architecture is an important aspect. From Table 2, per epoch train time is large for ViTs given they have on avg. around >70 M parameters to train.

On analysing, the results I found that we are not doing well to mitigate False positives, one reason for that could be class imbalance given pneumonia positive samples are large. Also we can do much better if we use pre-trained models and just train the fully connected layers instead of training full models on such a small dataset. From Table 2 and Table 1, the test times are within reasonable limits making ML/ DL techniques good for faster and accurate diagnosis of diseases on scale in the future.

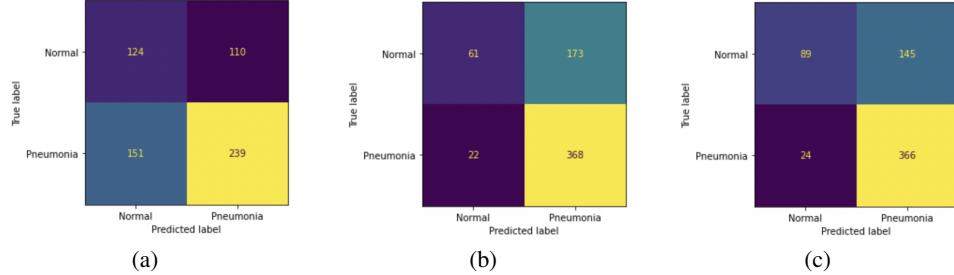


Figure 6: (a) PCA+Clustering (b) Logistic Regression (c) SVC

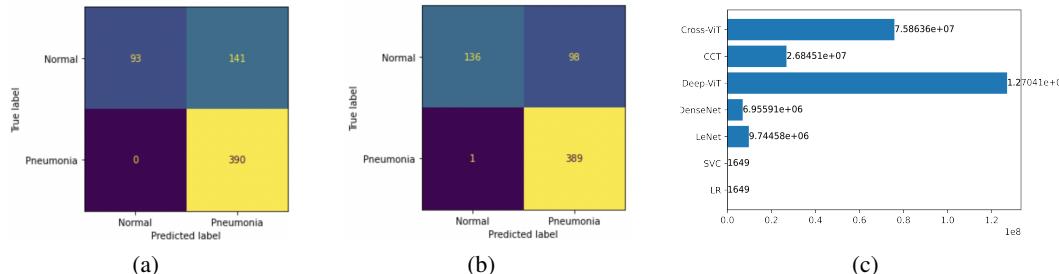


Figure 7: (a) Modified LeNet-5 (b) DenseNet-121 (c) No. of trainable parameters

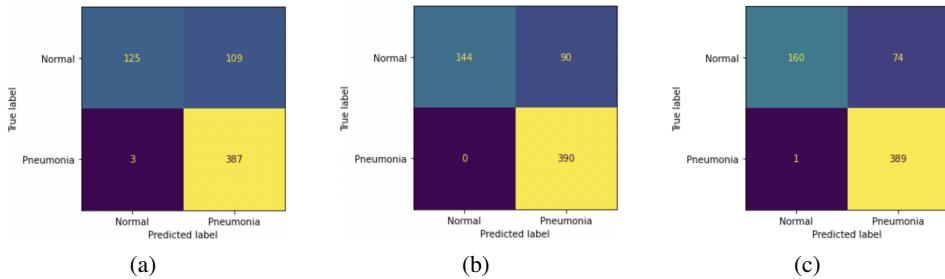


Figure 8: (a) Deep-ViT (b) CCT (c) Cross-ViT

Acknowledgements

I would like to thank the students of Machine Learning in Bioinformatics course, 2022 and Prof. Sriram Sankararaman for providing guidance for this project.

References

- [1] Gaurav Singh. *Pneumonia Detection*. <https://github.com/gauravSingh30/DL-PneumoniaDetection>. 28-Nov-2022.
- [2] World Health Organisation. *Pneumonia in children*. <https://www.who.int/news-room/fact-sheets/detail/pneumonia>. 11-Nov-2022.
- [3] Statista. *Number of deaths involving coronavirus disease*. <https://www.statista.com/statistics/1113051/number-reported-deaths-from-covid-pneumonia-and-flu-us/>. 25-Nov-2022.
- [4] Kermany et al. *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*. <https://data.mendeley.com/datasets/rscbjbr9sj/2>. 2018.
- [5] Paul Mooney. *Chest X-Ray Images (Pneumonia)*. <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>. 2017.
- [6] Irem Ersöz Kaya et al. “PCA based clustering for brain tumor segmentation of T1w MRI images”. In: *Computer methods and programs in biomedicine* 140 (2017), pp. 19–28.
- [7] Pranav Rajpurkar et al. “CheXnet: Radiologist-level pneumonia detection on chest x-rays with deep learning”. In: *arXiv preprint arXiv:1711.05225* (2017).
- [8] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [9] Daquan Zhou et al. “Deepvit: Towards deeper vision transformer”. In: *arXiv preprint arXiv:2103.11886* (2021).
- [10] Ali Hassani et al. “Escaping the big data paradigm with compact transformers”. In: *arXiv preprint arXiv:2104.05704* (2021).
- [11] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. “Crossvit: Cross-attention multi-scale vision transformer for image classification”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 357–366.