Bone X-Ray Classification for Upper Extremity Radiographs

Bhavesh Sood 2019335 bhavesh19355@iiitd.ac.in
Vishwajeet Kumar 2019128 vishwajeet19128@iiitd.ac.in
Abhimanyu Gupta 2019226 abhimanyu19226@iiitd.ac.in

Abstract

The aim is to detect abnormalities in upper extremity of human body, with better than human performance. One or more radiographs of a corresponding musculoskeletal study has been used as a feature set. The task is a binary classification, where each musculoskeletal study is classified as normal or abnormal. The training dataset consisted of 40,561 multi-view radiographic images from around 14 thousand studies, where each study is manually labelled by radiologists. After much experimentation and tweaking the models, we achieved best performance of 78.57% accuracy and 78.43% F1 Score, which is on par with the MURA dataset producer's kappa statistic of 0.778. The task in general is a good example to showcase the high usability of modern deep learning algorithms in the medical field providing great accuracy.

1 Introduction

In 2012, it was estimated that 1.7 billion people are affected by musculoskeletal conditions worldwide (Stuart I. Weinstein, 2014). These musculoskeletal conditions are a prominent reason for severe, long-term pain and disability. Moreover, a number of studies like (Krupinski and Kim, 2010), have shown how radiologists precision in their work tends to decrease due to the hectic workload they are exposed to. This posses a great threat to precious human life. With recent advents in Deep Learning techniques, several state of the art models have appeared which provides expert level performance on critical tasks like Detection of Diabetic Retinopathy (Gulshan, 2016). Thus, we anticipate a Deep Learning based model to perform the task of "Identifying Abnormalities in Musculoskeletal Radiographs" with comparable human accuracy (if not greater), saving precious human hours and lives.

Determining whether a radiographic study is normal or abnormal is a critical radiological task: a study interpreted as normal, rules out disease and can eliminate the need for patients to undergo further diagnostic procedures or interventions.

2 Problem Statement

The task is to classify a musculoskeletal radiograph study of the upper extremity ,where each study contains one or more views into normal or abnormal category with an accuracy of the level of a radiologist. We have a labelled dataset of 40,561 musculoskeletal radiograph images from over 14k studies of different parts of body provided by the MURA dataset collected by Picture Archive and Communication System (PACS) of Stanford Hospital. Each study contains set of images of a particular part of a patient from different angles. For instance suppose a patient x goes for X-Ray of his hand than the radiologist takes the radiograph from atleast 2 angles so these 2 images makes one study and we try to predict on the basis of those two images whether that part of patient is normal or it has some abnormalities in it, for e.g. fractured hand or lesions.

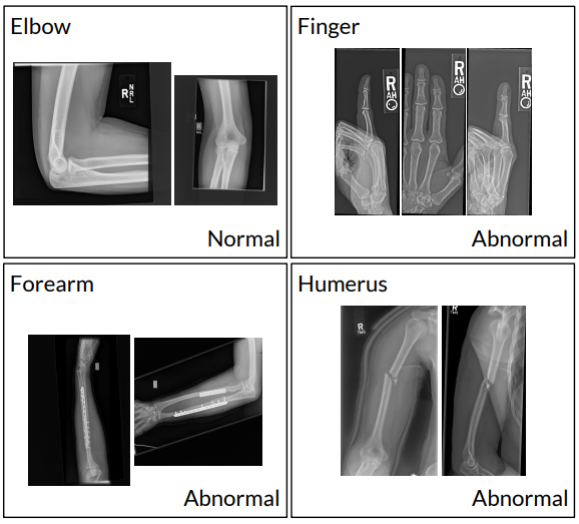

Figure 1: Different studies and their coresponding labels provided by radiologists

3 Dataset

The Mura dataset consists of 40,561 images from 14,863 studies. The dataset is made public by stanford university with their baseline model. The model contains images from 14k studies of approx 13k patients and it is labelled by the radiologists of those patients as normal and abnormal. We were able to find the data from this site : (stanford, 2018).

Study	Train		Validation		Total
	Normal	Abnormal	Normal	Abnormal	
Elbow	1094	660	92	66	1912
Finger	1280	655	92	83	2110
Hand	1497	521	101	66	2185
Humerus	321	271	68	67	727
Forearm	590	287	69	64	1010
Shoulder	1364	1457	99	95	3015
Wrist	2134	1326	140	97	3697
Total No. of Studies	8280	5177	661	538	14656

Figure 2: All regions and their coresponding number of studies

4 Related Work

4.1 MURA

Considering the importance of deep learning in mind, this paper contains the study for classsification of musculoskeletal radiographs into normal or abnormal using deep learning techniques. This paper introduces a dataset of 40k images over 14863 studies done by stanford university. They took help of several radiologists to prepare 207 more studies for test set and prposed a 169-layer densenet baseline model to train the data. The authors were able to achieve an AUROC of 0.929. The model performs really well for finger and wrist region. However their performance for elbow, forearm, hand, humerus and shoulder was not at par with that of finger and wrist region. (Rajpurkar, 2018).

4.2 DenseNet

In recent works we have seen that convolutions networks are becoming deeper and deeper to get more accurate. However as the input or the gradients pass through many layers, they can vanish or "wash-out" by the time they reach the end (or beginning) of the network. This is called the vanishing gradient problem . Many solutions have been proposed in the recent times such as ResNets (Kaiming He and Sun., 2015.) and Highway Networks (Rupesh K Srivastava and Schmidhuber., 2015.), all of which try to make shorter connections between layers close to the input and output. DenseNet embraces this observation by connecting each layer to every other layer in a feed-forward fashion. The traditional convolutional networks with L layers have L connections - one between each layer, in DenseNet each layer is connected to all layers after it , having $\frac{L(L+1)}{2}$ direct connections. Since there is no need to re-learn redundant feature maps it requires fewer parameters too. DenseNet obtains significant improvements over most state-of-the-art models. (Gao Huang and Weinberger., 2018).

4.3 ImageNet

ImageNet is a collection of majority of the 80,000 sysnets of WordNet structure with an average of 500-1000 clean and full resolution images of those words. This provides a way to index, retrieve, organize and interact with images and multimedia data. The current 12 subtrees consists of a total of 3.2 million clearly annotated images spread over 5247 categories. On an average over 600 images are collected for each sysnet. The MURA model initialised the weights for its model with the weights of a pretrained model on ImageNet. (Jia Deng and Li., 2009.).

4.4 ChestX-ray8

This paper presents a new chest X-Ray database and tries to classify the X-Ray into 8 diseases classes. The paper contains X-ray radiographs of people who were suffering from one of the following 8 diseases : Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumathorax and their labels based on the radiologists feedback. Then they try to train a model which can classify the images into above 8 categories. Their multi-label CNN architecture is implemented using caffe framework. They are using imageNet pre trained models i.e. AlexNet, GoogLeNet etc. Their model was able to recognize Cardiomegaly and Pneumothorax labels with an avg AUC of 0.8 while other labels had a low AUC of approx 0.6. This dataset and model are good benchmarks for further improving the fully-automated medical diagnosing systems. (Wang and Summers, 2017.)

4.5 Learning Deep Features for Discriminative Localization

The following paper presents a technique to localise the discriminative regions of images. A vital step in the classification problem is to identify the discriminative regions of the image. Once the discrimina-

tive regions are identified, the model can then use this information to classify the whole image into one of the given labels. Beside its utility in learning, this localisation also leads to better model interpretation by users. In the paper, Global Average Pooling (GAP) (with adequate tweeking), clubbed with a Class Activation Map (CAM) has been used to identify the prominent regions in a single forward pass. The authors of the (Rajpurkar, 2018), also uses the following technique to highlight the salient regions of radiographs which contribute the most to their abnormality detection prediction. (Zhou and Torralba, 2016)

4.6 Current Applications and Future Impact of Machine Learning in Radiology

The following paper walks through the possible (and currently used) application of ML/DL in the radiology. With likes of Detection and Interpretation of Radiology findings, the authors enlightens, the readers into the promises of ML/DL techniques and the future they could hold. Additionally, reviews of a number of currently used ML/DL techniques in the radiology is also presented. The current challenges of such techniques and expected future (along with possible extension) for them is also well discussed. (Garry Choy, 2018)

5 Methodology

Image was prepossessed by resizing them to 224 * 224, followed by normalization between -1 to 1. For training batches of 64 were made and sent to the model. The model consisted of DenseNet followed by a number of fully connected layers. Different weight initialisation techniques like, He and Xavier were also employed. Different combinations of Batch Normalisation, Dropout (varied layers, different probabilities) Optimiser and learning rate were also enforced.

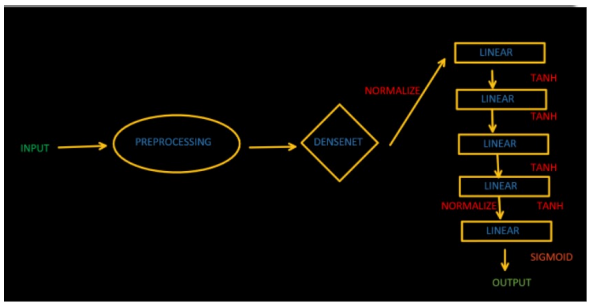

Figure 3: Model Overview (Simplified)

6 Experiment

A number of varied combinations of models were used for experimenting. Some models only had a single fully connected layer after DenseNet while some had up to 3 such layers. Further, to see the effect of freezing layers, again various combinations were made like freezing all the layers, freezing some, freezing only additional on top layers and so. To understand the role of weight initialisation techniques like He, Xavier and Random, different combinations were again made with fine tunning other parameters like learning rate, etc. Moreover, several combinations of Dropout and Batch Normalisation were also made. Overall, all the hyperparameters were mixed and different combinations were tested over different epochs to find the best model.

7 Results and Analysis

Model Configuration	Layers Freezed	Accuracy (%)	F1 Score (%)
Dropout (0.2) + BatchNorm	All	67.0	67.0
Dropout (0.2) + BatchNorm + He	None	77.6	77.6
Dropout (0.2) + BatchNorm + Xavier	None	58.0	43.0
Dropout (0.2) + Xavier + lr = 1e-4	Last 2 Denseblock Unfreeze	75.0	75.0
Dropout (0.2) + BatchNorm + He	Last 1 Denseblock Unfreeze	74.0	74.0
Dropout (0.2) + Random + lr = 1e-4	Last 3 Denseblock Unfreeze	77.1	77.0
More linear layers on top of DenseNet	None	78.5	78.5

Figure 4: Accuracy and F1 Score on various combinations

Our best model was able to achieve best performance of 78.57 % accuracy and 78.43 % F1 Score,

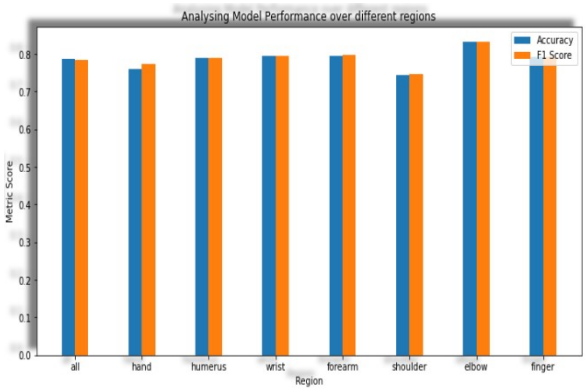

Figure 5: Best model Accuracy and F1 Score on different regions

which is on par with the MURA dataset producer's kappa statistic of 0.778. Furthermore, even on least performing regions the model accuracy was more than 70%. Comparing the best model with other models, we can see how adding more linear layers, i.e. making the model deep, was crucial for gaining performance. One reason we think might be the cause that freezing certain layers of DenseNet didn't increase performance, could be that the images we had were the x-ray images , but DenseNet was trained over ImageNet which contained general pictures and not the x-ray film like pictures.

8 Individual Contribution

- **Abhimanyu Gupta** : Performed region-wise analysis of model, Testing of the model with different optimizers, learning rates and layers.
- **Bhavesh Sood** : Loading the dataset and converting to pytorch dataloader in batches. Loading the pre-trained DenseNet. Fine-tuning some/all layers of the DenseNet model.
- **Vishwajeet Kumar** : Created Accuracy, fit plotted curves and visualisation of data, Testing of the model with different optimizers, learning rates and layers.

9 Future Work

There is always some scope for improvement in the models for such predictions , this can be achieved by trying different network architectures , varying the depth , tuning the hyperparameters etc. We can also try to categorize the abnormalities further into their types such as fractures, degenerative joint diseases, hardware , and other miscellaneous abnormalities (like lesions and subluxations).

We can try and train such deep learning models on other medical classifications and try to automate the medical diagnosing system, which can be very helpful to analyse a patient without any human interventions.

References

Laurens van der Maaten Gao Huang, Zhuang Liu and Kilian Q. Weinberger. 2018. Densely connected convolutional networks. arXiv:1608.06993.

Mark Michalski Synho Do Anthony E. Samir Oleg S. Pianykh J. Raymond Geis Pari V. Pandharipande James A. Brink Keith J. Dreyer. Garry Choy, Omid Khalilzadeh1. 2018. Current applications and future impact of machine learning in radiology. *Radiology 2018, 288*, page 318–328.

Peng Lily Coram Marc Stumpe Martin C Wu Derek Narayanaswamy Arunachalam Venugopalan Subhashini Widner Kasumi Madams Tom Cuadros Jorge et al. Gulshan, Varun. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama, 316(22)*, page 2402–2410.

Richard Socher Li-Jia Li Kai Li Jia Deng, Wei Dong and Fei-Fei Li. 2009. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 10.1109/CVPR.2009.5206848:248–255.

Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. 2015. Deep residual learning for image recognition. arXiv: 1512.03385.

Berbaum Kevin S Caldwell-Robert T Schartz Kevin M Krupinski, Elizabeth A and John. Kim. 2010. Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology, 7(9)*, page 698–704.

Irvin J. Bagul A. Ding-D. Duan T. Mehta H. Yang B. Zhu K. Laird D. Ball R. L. Langlotz C. Shpanskaya K. Lungren M. P. Ng A. Y. Rajpurkar, P. 2018. Mura: Large dataset for abnormality detection in musculoskeletal radiographs.

Klaus Greff Rupesh K Srivastava and Jürgen Schmidhuber. 2015. Training very deep networks. *Advances in Neural Information Processing Systems*, page volume 28.

stanford. 2018. Mura dataset. https://docs.activeloop.ai/datasets/mura-dataset.

Sylvia I. Watkins-Castillo. Stuart I. Weinstein, Edward H. Yelin. 2014. The big picture. *BMUS: The Burden of Musculoskeletal Diseases in the United States*.

Peng Yifan Lu Le Lu Zhiyong Bagheri-Mohammadhadi Wang, Xiaosong and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint*, arXiv:1705.02315.:248–255.

Khosla Aditya Lapedriza Agata Oliva Aude Zhou, Bolei and Antonio. Torralba. 2016. Learning deep features for discriminative localization. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 2921–2929.