A

PROJECT REPORT

ON

Classification and Localisation of Abnormality in Musculoskeletal Radiograph

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

in

ELECTRICAL ENGINEERING

Submitted By:

Sunayana Gupta (2015UEE1754)

Yashi Baldaniya (2015UEE1440)

Anudeep Reddy Katta (2015UEE1712)

Samidha Mridul Verma (2015UEE1434)

Supervised By:

Dr. Rajesh Kumar

Professor

DEPARTMENT OF ELECTRICAL ENGINEERING

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

May, 2019

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY

JAIPUR

CANDIDATE DECLARATION

We hereby submit the Project Report entitled "**Classification and Localisation of Abnormality in Musculoskeletal Radiograph**" in the **Department of Electrical Engineering**, **Malaviya National Institute of Technology Jaipur**. We have worked under the supervision of **Dr. Rajesh Kumar**, Professor and declare that the work presented in this report is an authentic record of our original work carried out under the guidance of our supervisor.

SUNAYANA GUPTA	**YASHI BALDANIYA**	**ANUDEEP REDDY**	**SAMIDHA VERMA**
2015UEE1754	2015UEE1440	2015UEE1712	2015UEE1434

The candidates have successfully completed the project under my supervision and the above statements made by them are true to the best of my knowledge. The Project Report is hereby approved for submission.

Date: 7 May 2019

Dr. Rajesh Kumar

Supervisor

ACKNOWLEDGEMENTS

It was a privilege to study and work under the guidance of Dr. Rajesh Kumar, Professor, Malaviya National Institute of Technology, Jaipur. We take this opportunity to express our deep sense of gratitude to him for his keen interest, valuable suggestions and constant encouragement throughout the project work. He made us comfortable and free to work due to his relaxed way of supervision.

We are thankful to the Ph.D. scholars for their support and needful guidance. Their suggestions and moral support helped us to complete our work. Last but not the least, we would like to thank our parents for their invaluable support, effort, and unconditional love.

Date : 7 May 2019
Place : MNIT, Jaipur

(Sunayana)

(Yashi)

(Anudeep)

(Samidha)

ABSTRACT

Musculoskeletal conditions affect more than 1.7 billion people worldwide, and are the most common cause of severe, long-term pain and disability, with 30 million emergency department visits annually and increasing. The MURA dataset can lead to significant advances in medical imaging technologies which can diagnose at the level of experts, towards improving healthcare access in parts of the world where access to skilled radiologists is limited. 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. Machine and Deep learning frameworks can be used for the classification of these radiographs as normal or abnormal images. Our project focuses on improvements on machine and deep learning frameworks for detection of abnormality in Musculoskeletal Radiographs.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full form
NasNet	Neural Architecture Search
CAM	Class Activation Mapping
ResNet	Residual Neural Network
PET	Positron Emission Tomography
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit
MURA	Musculoskeletal Radiographs

NOMENCLATURE

[x0, x1, . . . , x`−1]	Concatenation of the feature-maps produced in layers 0, . . . , −1
Hl(·)	as a composite function of three consecutive operations: batch normalization (BN), followed by a rectified linear unit (ReLU) and a 3 × 3 convolution (Conv)
k	Growth rate of densenet
l	Lth layer of densenet

CHAPTER 1
INTRODUCTION

Radiological imaging is of increasing importance in patient care. Both diagnostic and therapeutic indications for radiologic imaging are expanding rapidly. The rapid expansion is a consequence of the need for more rapid, accurate, cost-effective, and less invasive treatment. Technological advancements in radiologic imaging equipment have also fueled the utilization of imaging. Such technological advancements include the capability to acquire higher and higher resolution images, enabling visualization of smaller anatomic structures and abnormalities. The higher resolution comes at the cost of an ever-increasing average number of images per patient. Radiologists need to interpret these images and as the number of images increases, radiologists' workload increases as well. The increasing number and complexity of the images threaten to overwhelm radiologists' capacities to interpret them. In many real radiologic practices, automated and intelligent image analysis and understanding are becoming an essential part or procedure, such as image segmentation, registration, and computer-aided diagnosis and detection. In addition, in the area of cancer prognosis and treatment, automated and intelligent algorithms have a large market and are welcomed broadly, in areas such as radiation therapy planning or automatic identification of imaging biomarkers from radiological images of certain diseases, etc. Machine learning algorithms underpin the algorithms and software that make computer-aided diagnosis/prognosis/treatment possible.

Radiology is a branch of medical science which uses imaging technology and radiation to make diagnoses and treat disease. It has benefited greatly from the advances of physics, electronic engineering, and computer science. Based on different detection and imaging rationale, various modalities were developed in the past decades in the field of diagnostic radiology. Today, the mainstream modalities which are widely used in hospitals and medical centers include radiography, fluoroscopy, computed tomography (CT), ultrasound, magnetic resonance imaging (MRI), and positron emission tomography (PET).

In the daily practice of radiology, medical images from different modalities are read and interpreted by radiologists. Usually, radiologists must analyze and evaluate these images comprehensively in a short time. But with the advances in modern medical technologies, the amount of imaging data is rapidly increasing. For example, CT examinations are being performed with thinner slices than in the past. The reading and interpretation time of radiologists will mount as the number of CT slices grows.

Machine learning provides an effective way to automate the analysis and diagnosis of medical images. It can potentially reduce the burden on radiologists in the practice of radiology. The applications of machine learning in radiology include medical image segmentation (e.g., brain, spine, lung, liver, kidney, colon); medical image registration (e.g., organ image registration from different modalities or time series); computer-aided detection and diagnosis systems for CT or MRI images (e.g., mammography, CT colonography, and CT lung nodule CAD); brain function or activity analysis and neurological disease diagnosis from fMR images; content-based image retrieval systems for CT or MRI images; and text analysis of radiology reports using natural language processing (NLP) and natural language understanding (NLU).

Machine learning is the study of computer algorithms which can learn complex relationships or patterns from empirical data and make accurate decisions. It is an interdisciplinary field that has close relationships with artificial intelligence, pattern recognition, data mining, statistics, probability theory, optimization, statistical physics, and theoretical computer science. Applications of machine learning include natural language processing, medical diagnosis, bioinformatics, video surveillance, and financial data analysis.

Machine learning algorithms can be organized into different categories based on different principles. For example, depending on the utilization of labels of training samples, they can be categorized into supervised learning, semi-supervised learning, and unsupervised learning algorithms.

In supervised learning, each sample contains two parts: one is input observations or features and the other is output observations or labels. Usually, the input observations are causes and the output observations are effects. The purpose of supervised learning is to deduce a functional

relationship from training data that generalizes well to testing data. The form of the relationship is a set of equations and numerical coefficients or weights. Examples of supervised learning include classification, regression, and reinforcement learning.

In unsupervised learning, we only have one set of observations and there is no label information for each sample. Usually, these observations or features are caused by a set of unobserved or latent variables. The main purpose of unsupervised learning is to discover relationships between samples or reveal the latent variables behind the observations. Examples of unsupervised learning include clustering, density estimation, and blind source separation.

Semi-supervised learning falls between supervised and unsupervised learning. It utilizes both labeled data (usually a few) and unlabeled data (usually many) during the training process. Semi-supervised learning algorithms were developed mainly because the labeling of data is very expensive or impossible in some applications. Examples of semi-supervised learning include semi-supervised classification and information recommendation systems.

Chapter 1 is concerned with radiological imaging and its importance, basic overview of machine learning including supervised, unsupervised and semi-supervised techniques and how it can be an effective method for analysis and diagnosis of medical images.

Chapter 2 gives a detailed account of the literature survey, which involves various works of collection large medical datasets and applying deep learning algorithms to achieve human-level performance on tasks such as image recognition, like detection of abnormalities, tumors or diseases, traditionally done by highly trained radiologists.

Chapter 3 gives a brief description about the dataset, on how it was collected and can be used for the purpose of detecting and classifying abnormalities in musculoskeletal radiographs.

Chapter 4 focuses on the problem formulation on why do we need such a robust system for detection of musculoskeletal abnormalities. It also addresses the problem with radiologists and diagnosing these images manually. These problems can be fixed by creating a robust deep learning architecture.

Chapter 5 starts with data pre-processing, one of the key processes in the procedure of detecting abnormalities and diagnosis of medical images. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues and this section presents techniques involved in it.

Chapter 6 describes the different deep learning models that have been implemented and are explained in detail. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

Chapter 7 and 8 elaborate the results and conclusions obtained from implementing models mentioned in Chapter 6.

CHAPTER 2
LITERATURE REVIEW

Deep learning algorithms, powered by advances in computation and very large datasets, have recently been shown to exceed human performance in visual tasks such as playing Atari games, strategic board games like Go and object recognition. In the era of deep learning in computer vision, research efforts on building various annotated image datasets with different characteristics play indispensably important roles on the better definition of the forthcoming problems, challenges and subsequently possible technological progresses. Particularly, here we focus on the relationship and joint learning of image (radiographs) and text (radiograph reports). The previous representative image caption generation work utilize Flickr8K, Flickr30K and MS COCO datasets that hold 8,000, 31,000 and 123,000 images respectively and every image is annotated by five sentences via Amazon Mechanical Turk (AMT). The text generally describes annotator's attention of objects and activity occurring on an image in a straightforward manner. Region-level ImageNet pre-trained convolutional neural networks (CNN) based detectors are used to parse an input image and output a list of attributes or "visually-grounded high-level concepts" (including objects, actions, scenes and so on). Visual question answering (VQA) requires more detailed parsing and complex reasoning on the image contents to answer the paired natural language questions. A new dataset containing 250k natural images, 760k questions and 10M text answers is provided to address this new challenge. Additionally, databases such as "Flickr30k Entities", "Visual7W" and "Visual Genome" (as detailed as 94,000 images and 4,100,000 region-grounded captions) are introduced to construct and learn the spatially-dense and increasingly difficult semantic links between textual descriptions and image regions through the object-level grounding.

While previous classifiers mostly used shallow neural networks, recent years witnessed great advancement on applying deep learning to computer aided detection. Wang et al. introduced ChestX-ray8, a hospital-scale chest X-ray database, and provided benchmarks on weakly-supervised classification and localization of common thorax diseases. They applied deep CNNs and added transition layers to produce heatmap for localization. Following this work,

Rajpurkar et al. introduced CheXNet, a 121-layer Dense Convolutional Network (DenseNet) trained on the Chest X-ray 14 dataset, producing radiologist-level pneumonia detection. Moreover, Rajpurkar et al. introduced MURA dataset for detecting radiologist level abnormality in musculoskeletal radiographs. Deep learning has also been applied to segmenting important structures from medical images. Fu et al. applied multi-scale and multi-level CNN with a side-output layer to learn hierarchical representation of features. To model the long-range interactions between pixels, Conditional Random Field (CRF) is used.

Machine learning has been applied to medical measurements and imaging applications. Rosati et al. used multiparametric MRI along with a clustering procedure based on self-organizing map (SOM) to improve the detection of prostate cancer. Roza et al. presented an artificial neural network (ANN) to identify two types of arrhythmias in ECG signals. Alkabawi et al. proposed an approach for computer-aided classification of multi-types of dementia using convolutional neural networks.

Anatomical object localization (in space or time), such as organs or landmarks, has been an important pre-processing step in segmentation tasks or in the clinical workflow for therapy planning and intervention. Localization in medical imaging often require sparsing of 3D volumes. To solve 3D data parsing with deep learning algorithms, several approaches have been proposed that treat the 3D space as a composition of 2D orthogonal planes. Yang et al. (2015) identified landmarks on the distal femur surface by processing three independent sets of 2D MRI slices (one for each plane) with regular CNNs. Other authors try to modify the network learning process to directly predict locations. For example, Payer et al. (2016) proposed to directly regress landmark locations with CNNs. They used landmark maps, where each landmark is represented by a Gaussian, as ground truth input data and the network is directly trained to predict this landmark map. Another interesting approach was published by Ghesu et al. (2016), in which reinforcement learning is applied to the identification of landmarks. The authors showed promising results in several tasks: 2D cardiac MRI and ultrasound (US) and 3D head/neck CT.

Due to its increased complexity, only a few methods addressed the direct localization of landmarks and regions in the 3D image space. Zheng et al. (2015) reduced this complexity by

decomposing 3D convolution as three one-dimensional convolutions for carotid artery bifurcation detection in CT data. Ghesu et al. (2016) proposed as parse adaptive deep neural network powered by marginal space learning in order to deal with data complexity in the detection of the aortic valve in 3D transesophageal echocardiogram.

CNNs have also been used for the localization of scan planes or key frames in temporal data. Baumgartner et al. (2016) trained CNNs on video frame data to detect upto 12 standardized scan planes in mid-pregnancy fetal US. Furthermore, they used saliency maps to obtain a rough localization of the object of interest in the scan plan (e.g. brain, spine). RNNs, particularly LSTM-RNNs, have also been used to exploit the temporal information contained in medical videos, another type of high dimensional data. Chen et al. (2015), for example, employed LSTM models to incorporate temporal information of consecutive sequence in US videos for fetal standard plane detection. Kong et al. (2016) combined an LSTM-RNN with a CNN to detect the end-diastole and end-systole frames in cine-MRI of the heart.

Concluding, localization through 2D image classification with CNNs seems to be the most popular strategy over all to identify organs, regions and landmarks, with good results. However, several recent papers expand on this concept by modifying the learning process such that accurate localization is directly emphasized, with promising results.

The detection of objects of interest or lesions in images is a key part of diagnosis and is one of the most labor-intensive for clinicians. Typically, the tasks consist of the localization and identification of small lesions in the full image space. There has been a long research tradition in computer-aided detection systems that are designed to automatically detect lesions, improving the detection accuracy or decreasing the reading time of human experts. Interestingly, the first object detection system using CNNs was already proposed in 1995, using a CNN with four layers to detect nodules in x-ray images (Lo et al., 1995). Most of the published deep learning object detection systems still uses CNNs to perform pixel (or voxel) classification, after which some form of post processing is applied to obtain object candidates.

The incorporation of contextual or 3D information is also handled using multi-stream CNNs (for example by Barbu et al., 2016 and Roth et al., 2016). Teramoto et al. (2016) used a multi-stream

CNN to integrate CT and Positron Emission Tomography (PET) data. Dou et al. (2016b) used a 3D CNN to find micro-bleeds in brain MRI. Last, as the annotation burden to generate training data can be similarly significant compared to object classification, weakly-supervised deep learning has been explored by Hwang and Kim (2016), who adopted such a strategy for the detection of nodules in chest radiographs and lesions in mammography.

The segmentation of organs and other substructures in medical images allows quantitative analysis of clinical parameters related to volume and shape, as, for example, in cardiac or brain analysis. Furthermore, it is often an important first step in computer-aided detection pipelines. The task of segmentation is typically defined as identifying the set of voxels which make up either the con- tour or the interior of the object(s) of interest. Segmentation is the most common subject of papers applying deep learning to medical imaging, and as such has also seen the widest variety in methodology, including the development of unique CNN-based segmentation architectures and the wider application of RNNs.

The most well-known, in medical image analysis, of these novel CNN architectures is U-net, published by Ronneberger et al. (2015). The two main architectural novelties in U-net are the combination of an equal amount of upsampling and downsampling layers. Although learned upsampling layers have been proposed before, U-net combines them with so-called skip connections between opposing convolution and deconvolution layers. This which concatenate features from the contracting and expanding paths. From a training perspective this means that entire images/scans can be processed by U-net in one forward pass, resulting in a segmentation map directly. This allows U-net to take into account the full context of the image, which can be an advantage in contrast to patch-based CNNs. Furthermore, in an extended paper by Çiçek et al. (2016), it is shown that a full 3D segmentation can be achieved by feeding U-net with a few 2D annotated slices from the same volume. Other authors have also built derivatives of the U-net architecture; Milletari et al. (2016b), for example, proposed a 3D-variant of U-net architecture, called V-net, performing 3D image segmentation using 3D convolutional layers with an objective function directly based on the Dice coefficient. Drozdzal et al. (2016) investigated the

use of short ResNet-like skip connections in addition to the long skip-connections in a regular U-net.

RNNs have recently become more popular for segmentation tasks. For example, Xie et al. (2016) used a spatial clockwork RNN to segment the perimysium in H&E-histopathology images. This network takes into account prior information from both the row and column predecessors of the current patch. To incorporate bidirectional information from both left/top and right/bottom neighbors, the RNN is applied four times in different orientations and the end-result is concatenated and fed to a fully-connected layer. This produces the final output for a single patch. Stollenga et al. (2015) where the first to use a 3D LSTM-RNN with convolutional layers in six directions. Andermatt et al. (2016) used a 3D RNN with gated recurrent units to segment gray and white matter in a brain MRI data set. Chen et al. (2016) combined bi-directional LSTM-RNNs with 2D U-net-like-architectures to segment structures in anisotropic 3D electron microscopy images. Last, Poudel et al. (2016) combined a 2D U-net architecture with a gated recurrent unit to perform 3D segmentation.

fCNNs have also been extended to 3D and have been applied to multiple targets at once: Korez et al. (2016), used 3D fC- NNs to generate vertebral body likelihood maps which drove deformable models for vertebral body segmentation in MR images, Zhou et al. (2016) segmented nineteen targets in the human torso, and Moeskops et al. (2016) trained a single fCNN to segment brain MRI, the pectoral muscle in breast MRI, and the coronary arteries in cardiac CT angiography (CTA).

DNNs have been extensively used for brain image analysis in several different application domains. A large number of studies address classification of Alzheimer's disease and segmentation of brain tissue and anatomical structures (e.g. the hippocampus). Other important areas are detection and segmentation of lesions (e.g. tumors, white matter lesions, lacunes, micro-bleeds). Apart from the methods that aim for a scan-level classification (e.g. Alzheimer diagnosis), most methods learn mappings from local patches to representations and subsequently from representations to labels. However, the local patches might lack the contextual information required for tasks where anatomical information is paramount (e.g. white matter lesion

segmentation). To tackle this, Ghafoorian et al. (2016) used non-uniformly sampled patches by gradually lowering sampling rate in patch sides to span a larger context. An alternative strategy used by many groups is multi-scale analysis and a fusion of representations in a fully-connected layer. Even though brain images are 3D volumes in all surveyed studies, most methods work in 2D, analyzing the 3D volumes slice-by-slice. This is often motivated by either the reduced computational requirements or the thick slices relative to in-plane resolution in some data sets. More recent publications had also employed 3D networks. DNNs have completely taken over many brain image analysis challenges. In the 2014 and 2015 brain tumor segmentation challenges (BRATS), the 2015 longitudinal multiple sclerosis lesion segmentation challenge, the 2015 ischemic stroke lesion segmentation challenge (ISLES), and the 2013 MR brain image segmentation challenge (MRBrains), the top ranking teams to date have all used CNNs. Almost all of the aforementioned methods are concentrating on brain MR images.

In thoracic image analysis of both radiography and computed tomography, the detection, characterization, and classification of nodules is the most commonly addressed application. Many works add features derived from deep networks to existing feature sets or compare CNNs with classical machine learning approaches using handcrafted features. In chest X-ray, several groups detect multiple diseases with a single system. In CT the detection of textural patterns indicative of interstitial lung diseases is also a popular research topic.

Chest radiography is the most common radiological exam; several works use a large set of images with text reports to train systems that combine CNNs for image analysis and RNNs for text analysis. This is a branch of research we expect to see more of in the near future. In a recent challenge for nodule detection in CT, LUNA16, CNN architectures were used by all top performing systems. This is in contrast with a previous lung nodule detection challenge, ANODE09, where handcrafted features were used to classify nodule candidates. The best systems in LUNA16 still rely on nodule candidates computed by rule-based image processing, but systems that use deep networks for candidate detection also performed very well (e.g. U-net). Estimating the probability that an individual has lung cancer from a CT scan is an important topic: It is the

objective of the Kaggle Data Science Bowl 2017, with $1 million in prizes and more than one thousand participating teams.

Deep learning techniques have also been applied for normalization of histopathology images. Color normalization is an important research area in histopathology image analysis. In Janowczyk et al. (2017), a method for stain normalization of hematoxylin and eosin (H&E) stained histopathology images was presented based on deep sparse auto-encoders. Recently, the importance of color normalization was demonstrated by Sethi et al. (2016) for CNN based tissue classification in H&E stained images. The introduction of grand challenges in digital pathology has fostered the development of computerized digital pathology techniques. The challenges that evaluated existing and new approaches for analysis of digital pathology images are: EM segmentation challenge 2012 for the 2D segmentation of neuronal processes, mitosis detection challenges in ICPR 2012 and AMIDA 2013, GLAS for gland segmentation and, CAMELYON16 and TUPAC for processing breast cancer tissue samples.

Musculoskeletal images have also been analyzed by deep learning algorithms for segmentation and identification of bone, joint, and associated soft tissue abnormalities in diverse imaging modalities. A surprising number of complete applications with promising results are available; one that stands out is Jamaludin et al. (2016) who trained their system with 12K discs and claimed near-human performances across four different radiological scoring tasks.

Deep learning has been applied to many aspects of cardiac image analysis. MRI is the most researched modality and left ventricle segmentation the most common task, but the number of applications is highly diverse: segmentation, tracking, slice classification, image quality assessment, automated calcium scoring and coronary centerline tracking, and super-resolution. Most papers used simple 2D CNNs and analyzed the 3D and often 4D data slice by slice; the exception is Wolterink et al. (2016) where 3D CNNs were used. DBNs are used in four papers, but these all originated from the same author group. The DBNs are only used for feature extraction and are integrated in com- pound segmentation frameworks. Two papers are exceptional because they combined CNNs with RNNs: Poudel et al. (2016) introduced a recurrent connection within the U-net architecture to segment the left ventricle slice by slice and

learn what information to remember from the previous slices when segmenting the next one. Kong et al. (2016) used an architecture with a standard 2D CNN and an LSTM to perform temporal regression to identify specific frames and a cardiac sequence.

Most papers on the abdomen aimed to localize and segment organs, mainly the liver, kidneys, bladder, and pancreas. Two papers address liver tumor segmentation. The main modality is MRI for prostate analysis and CT for all other organs. The colon is the only area where various applications were addressed, but always in a straightforward manner: a CNN was used as a feature extractor and these features were used for classification. It is interesting to note that in two segmentation challenges –SLIVER07 for liver and PROMISE12 for prostate –more traditional image analysis methods were dominant up until 2016. In PROMISE12, the current second and third in rank among the automatic methods used active appearance models. The algorithm from IMorphics was ranked first for almost five years (now ranked second). However, a 3D fCNN similar to U-net (Yu et al., 2017) has recently taken the top position. This paper has an interesting approach where a sum-operation was used instead of the concatenation operation used in U-net, making it a hybrid between a ResNet and U-net architecture. Also in SLIVER07 –a 10-year-old liver segmentation challenge –CNNs have started to appear in 2016 at the top of the leaderboard, replacing previously dominant methods focused on shape and appearance modeling.

It is clear that applying deep learning algorithms to medical image analysis presents several unique challenges. The lack of large training data sets is often mentioned as an obstacle. However, this notion is only partially correct. The use of PACS systems in radiology has been routine in most western hospitals for at least a decade and these are filled with millions of images. There are few other domains where this magnitude of imaging data, acquired for specific purposes, are digitally available in well-structured archives. PACS-like systems are not as broadly used for other specialties in medicine, like ophthalmology and pathology, but this is changing as imaging becomes more prevalent across disciplines. We are also seeing that increasingly large public data sets are made available: Esteva et al. (2017) used 18 public data sets and more than 10^5 training images; in the Kaggle diabetic retinopathy competition a similar number of retinal images were released; and several chest x-ray studies used more than 10^4

images. The main challenge is thus not the availability of image data itself, but the acquisition of relevant annotations/labeling for these images. Traditionally PACS systems store free-text reports by radiologists describing their findings. Turning these reports into accu- rate annotations or structured labels in an automated manner requires sophisticated text-mining methods, which is an important field of study in itself where deep learning is also widely used nowadays. With the introduction of structured reporting into several areas of medicine, distilling labels from these reports is expected to become easier in the future. For example, there are already papers appearing which directly leverage BI-RADS categorizations by radiologist to train deep networks (Kisilev et al., 2016) or semantic descriptions in analyzing optical coherence tomography images (Schlegl et al., 2015).

CHAPTER 3
DATASET

Large, high-quality datasets have played a critical role in driving the progress of fields with deep learning methods. MURA (**Mu**sculoskeletal **Ra**diographs), a large dataset of radiographs, containing 40,000 musculoskeletal images of the upper extremity. Each image is manually labeled by radiologists as either normal or abnormal. Our dataset, MURA, contains 9,045 normal and 5,818 abnormal musculoskeletal radiographic studies (each study has images from different views) of the upper extremity including the shoulder, humerus, elbow, forearm, wrist, hand, and finger.

The institutional review board approved study collected de-identified, HIPAA-compliant images from the Picture Archive and Communication System (PACS) of Stanford Hospital. They (the researchers given in the paper) assembled a dataset of musculoskeletal radiographs consisting of 14,863 studies from 12,173 patients, with a total of 40,561 multi-view radiographic images. Each belongs to one of seven standard upper extremity radiographic study types: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each study was manually labeled as normal or abnormal by board-certified radiologists from the Stanford Hospital at the time of clinical radiographic interpretation in the diagnostic radiology environment between 2001 and 2012. The labeling was performed during interpretation on DICOM images presented on at least 3 megapixel PACS medical grade display with max luminance 400 cd/m^2 and min luminance 1 cd/m^2 with a pixel size of 0.2 and a native resolution of 1500 x 2000 pixels. The clinical images vary in resolution and in aspect ratios. The dataset is split into training (11,184 patients, 13,457 studies, 36,808 images), validation (783 patients, 1,199 studies, 3,197 images), and test (206 patients, 207 studies, 556 images) sets. There is no overlap in patients between any of the sets.

To investigate the types of abnormalities present in the dataset, they reviewed the radiologist reports to manually label 100 abnormal studies with the abnormality finding: 53 studies were

labeled with fractures, 48 with hardware, 35 with degenerative joint diseases, and 29 with other miscellaneous abnormalities, including lesions and subluxations.

Table 1 summarizes the distribution of normal and abnormal studies.

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

Table 1: Table showing distribution of images in MURA dataset

Musculoskeletal conditions affect more than 1.7 billion people worldwide, and are the most common cause of severe, long-term pain and disability, with 30 million emergency department visits annually and increasing. The MURA dataset can lead to significant advances in medical imaging technologies which can diagnose at the level of experts, towards improving healthcare access in parts of the world where access to skilled radiologists is limited. MURA is one of the largest public radiographic image datasets. It consists of 40,561 images,where each study is manually labeled by radiologists as either normal or abnormal. Out of which 9,045 normal and 5,818 abnormal musculoskeletal radiographic studies of the upper extremity including the shoulder, humerus, elbow, forearm, wrist, hand, and finger.

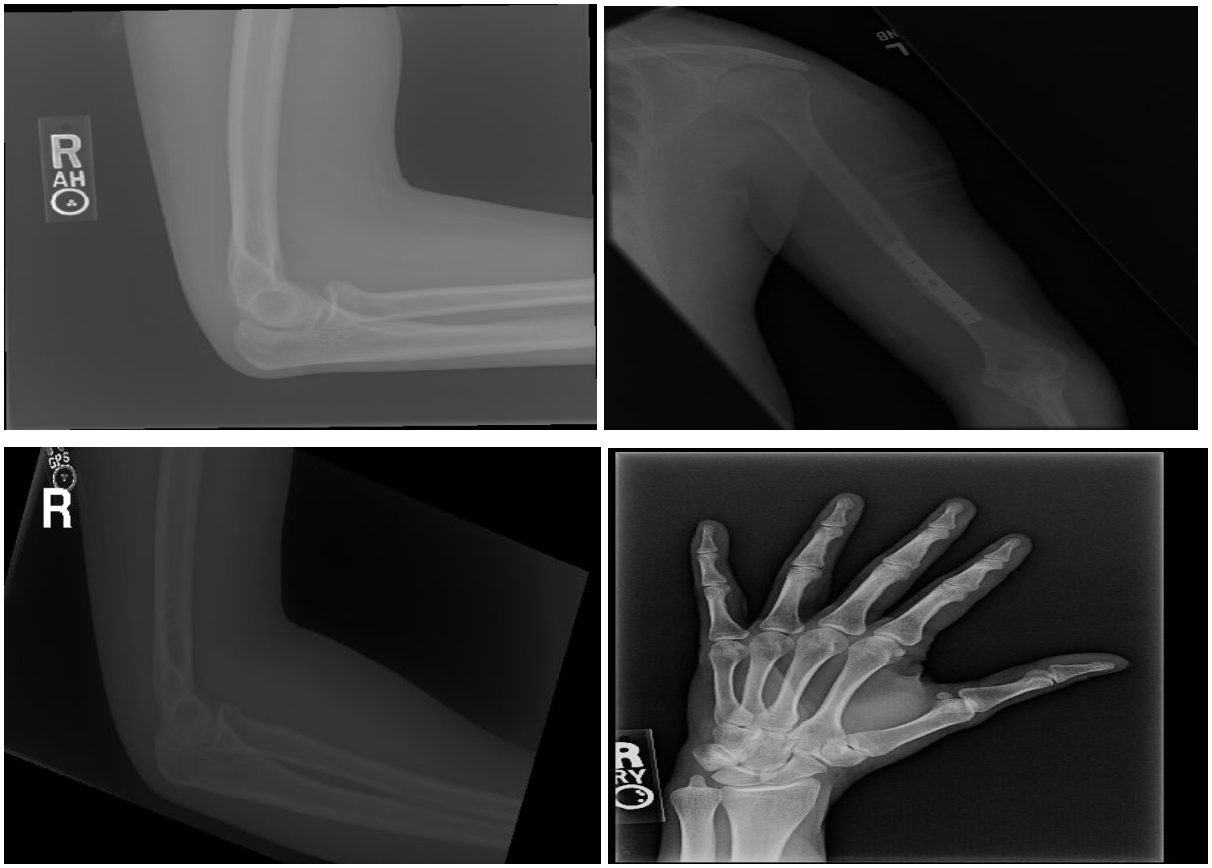


Fig 1: Sample images from MURA dataset

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.

CHAPTER 4
PROBLEM FORMULATION

4.1 The Burden of Musculoskeletal Diseases

Musculoskeletal conditions are among the most disabling and costly conditions suffered by people around the world. Musculoskeletal disorders and diseases are the leading cause of physical disability in our country. In December 2012, a study on the Global Burden of Disease and the worldwide impact of all diseases and risk factors found musculoskeletal conditions such as arthritis and back pain affect more than 1.7 billion people worldwide, are the second greatest cause of disability, and have the 4th greatest impact on the overall health of the world population when considering both death and disability. Professor Christopher Murray, lead investigator, and the authors of the study underline the need to address the rising numbers of individuals with a range of conditions such as musculoskeletal disorders that largely address disability, not mortality, in the future.

Bone and joint disorders account for more than one-half of all chronic conditions in people older than 50 years of age in developed countries, and are the most common cause of severe, long-term pain and disability. In spite of the widespread prevalence of musculoskeletal conditions and three of the most costly healthcare conditions—trauma, back pain, and arthritis—being musculoskeletal, musculoskeletal conditions are not among the top ten health conditions receiving research funding, primarily due to the low mortality from musculoskeletal conditions in comparison with other health conditions. However, the morbidity cost of musculoskeletal conditions is tremendous because musculoskeletal conditions often restrict activities of daily living, cause lost work days, and are a source of lifelong pain.

"Time and again, when the global burdens of disease are enumerated, musculoskeletal conditions rank high. Now we see that that rank is increasing. Although research funding reflects a long-term bias towards diseases with high mortality rates, the Global Burden of Disease project

indicates that much of the growth in disease burdens has occurred for conditions that cause high disability rates. Redressing the funding disparity should become a high priority."

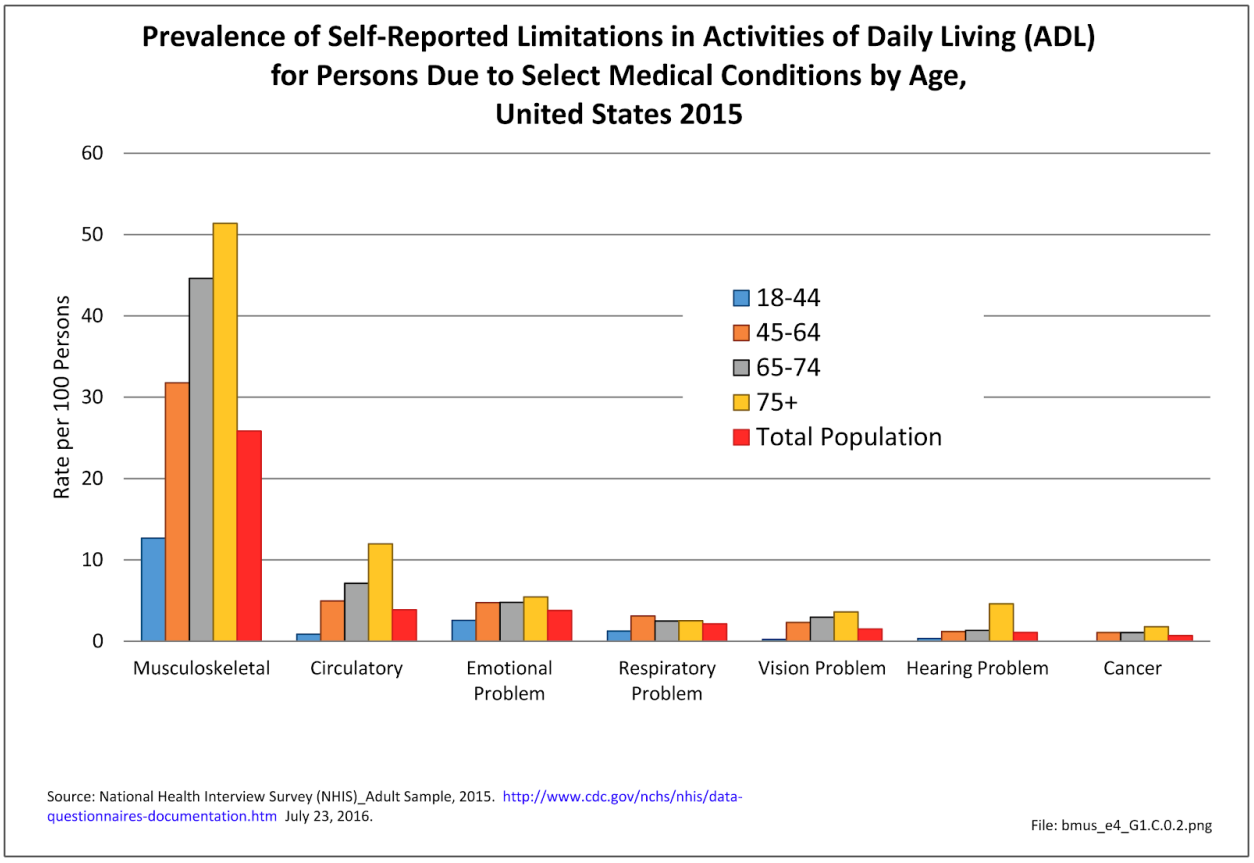


Fig 2: Prevalence of self-reported limitations in daily-living activities for people due to medical conditions by age

The rate of chronic musculoskeletal conditions found in the adult population is 76% greater than that of chronic circulatory conditions, which include coronary and heart conditions, and nearly twice that of all chronic respiratory conditions. On an age-adjusted basis, musculoskeletal conditions are reported by 54 persons per every 100 in the population.

Musculoskeletal conditions are found among all age groups, with the proportion of persons reporting these conditions increasing with age. Musculoskeletal conditions are reported by nearly three of four (70%) persons age 65 years and over. This compares to the 61% of persons age 65 to 74 years, and only slightly less than the 72% of those aged 75 years and older, reporting circulatory conditions, the majority of whom report chronic hypertension.

4.2 Liability of reading too many radiographs (Case Study):

A 54-year-old woman underwent screening mammography. The radiologist who later that day interpreted the mammograms immediately noted a 1.5-cm spiculated mass in the inferior inner portion of the left breast near the chest wall. Realizing that the lesion appeared highly suggestive of carcinoma, the radiologist reviewed the previous mammography study that had been obtained 1 year before, which he, himself, had interpreted as showing normal findings. In retrospect, the radiologist could now see the lesion on the mediolateral view, although only the anterior portion of the lesion was visible on the radiograph. The radiologist could not find the lesion on the craniocaudal view, but he immediately recognized that the breast had not been adequately positioned; the posterior part of the breast near the chest wall had not been shown on the radiograph. The radiologist rendered his written interpretation, describing the lesion, suggesting prompt biopsy, and adding that the lesion had been present although not reported on the previous study.

Biopsy of the breast lesion revealed infiltrating ductal carcinoma. The patient underwent mastectomy. All sampled lymph nodes were negative for tumor. Eight months later, the patient filed a medical malpractice lawsuit against the radiologist, alleging that the radiologist's failure to diagnose carcinoma on the initial mammography led to a 1-year delay in diagnosis that "substantially reduced the patient's chance for cure and normal life span." In addition to asking for compensatory damages, the attorney for the plaintiff, in an unusual move, also asked the court to award punitive damages because the "defendant radiologist read too many x-ray examinations on the day in question, demonstrating a wanton disregard of patient well-being by sacrificing quality patient care for volume in order to maximize revenue."

Allegations made in this case—that the radiologist missed a breast lesion on mammography because he was "overworked" by interpreting too many radiographic studies in a given day—are extremely unusual in medical malpractice lawsuits. Nonetheless, the allegations were taken seriously enough by both sides to effect a larger settlement than what might otherwise have resulted, given the bare facts of the case. A published study showing that the average radiologist's workload of 50-60 procedures a day was used by an expert witness for the plaintiff

to demonstrate that the defendant radiologist who admitted to interpreting 162 cases a day was indeed overworked and therefore negligent to be sure, and possibly reckless as well.

There are many such cases where the proper detection of abnormalities in the bones could have saved a life therefore we are proposing a solution which improves the prevalent situation by helping the radiologists in detection and localisation of bone abnormality. This can act as an aid to the radiologists while reading the radiographs.

CHAPTER 5
DATA PREPROCESSING

5.1 Data Augmentation

Invariance is the ability of convolutional neural networks to classify objects even when they are placed in different orientations. Data augmentation is a way of creating new 'data' with different orientations. The benefits of this are two-fold, the first being the ability to generate 'more data' from limited data and secondly it prevents overfitting. Our optimization goal is to chase that sweet spot where our model's loss is low, which happens when the parameters are tuned in the right way. If there are a lot of parameters, we would need to show machine learning model a proportional amount of examples, to get good performance. Also, the number of parameters needed is proportional to the complexity of the task model has to perform. More specifically, a CNN can be invariant to translation, viewpoint, size or illumination (Or a combination of the above). This essentially is the premise of data augmentation. In the real world scenario, we may have a dataset of images taken in a limited set of conditions. But, our target application may exist in a variety of conditions, such as different orientation, location, scale, brightness etc. We account for these situations by training our neural network with additional synthetically modified data i.e. data after augmentation.

Here, we are performing following data augmentation techniques:

(i) Resizing of images -

The images which are input ot the machine learning model should be of the same size and this resizing of images is very much vital task to be carried out.

(ii) Horizontal Flip -

An image flip means reversing the rows or columns of pixels in the case of a vertical or horizontal flip respectively.

(iii) Crop from Center -

Crops the image from centre discarding the irrelevant features of the image and selecting only the relevant features.

(iv) Random rotation of images -

A rotation augmentation randomly rotates the image clockwise by a given number of degrees from 0 to 360.

The rotation will likely rotate pixels out of the image frame and leave areas of the frame with no pixel data that must be filled in.

(v) Sharpening of images -

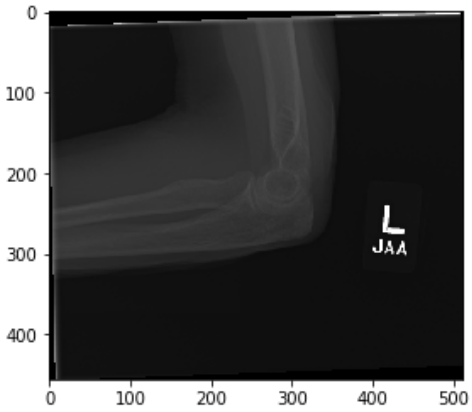
Sharpening of images highlights the boundary/lines creating more contrast for the area of interest and background.

(vi) Gaussian Blurring of images -

Gaussian blur (also known as Gaussian smoothing) is the result of blurring an image by a Gaussian function. It is used to reduce image noise.

5.2 Data Normalisation

Histogram equalization is a method to process images in order to adjust the contrast of an image by modifying the intensity distribution of the histogram. The idea of this processing is to give to the resulting image a linear cumulative distribution function. The local contrast of the object in the image is increased by applied histogram equalization, especially when the applied data of the image is represented by close contrast values. Through this adjustment the intensity can be better distributed on the histogram, this allows for areas of lower local contrast to gain a higher contrast without affecting the global contrast.



(a)

(b)

Fig 3: (a) Before Preprocessing (b) After Preprocessing

CHAPTER 6
METHODOLOGY

We have implemented the different learning models which are explained below in detail:

6.1 Convolutional Neural Network (CNN):

2012 was the first year that neural nets grew to prominence as Alex Krizhevsky used them to win that year's ImageNet competition (basically, the annual Olympics of computer vision), dropping the classification error record from 26% to 15%, an astounding improvement at the time. Ever since then, a host of companies have been using deep learning at the core of their services. Facebook uses neural nets for their automatic tagging algorithms, Google for their photo search, Amazon for their product recommendations, Pinterest for their home feed personalization, and Instagram for their search infrastructure.

CNN is popularly used for image processing and particularly for image classification.

Convolution: The first layer in a CNN is always a Convolutional Layer. Let the input be a 6 x 6 x 3 array of pixel values. Now, the best way to explain a conv layer is to imagine a flashlight that is shining over the top left of the image. Let's say that the light this flashlight shines covers a 3 x 3 area. And now, let's imagine this flashlight sliding across all the areas of the input image. In machine learning terms, this flashlight is called a filter(or sometimes referred to as a neuron or a kernel) and the region that it is shining over is called the receptive field. Now this filter is also an array of numbers (the numbers are called weights or parameters). A very important note is that the depth of this filter has to be the same as the depth of the input (this makes sure that the math works out), so the dimensions of this filter is 3 x 3 x 3. As the filter is sliding, or convolving, around the input image, it is multiplying the values in the filter with the original pixel values of the image (aka computing element wise multiplications). These multiplications are all summed up (mathematically speaking, this would be 27 multiplications in total). Every unique location on the input volume produces a number. After sliding the filter over all the locations, a 4 x 4 x 1 array of numbers, an activation map or feature map is obtained.

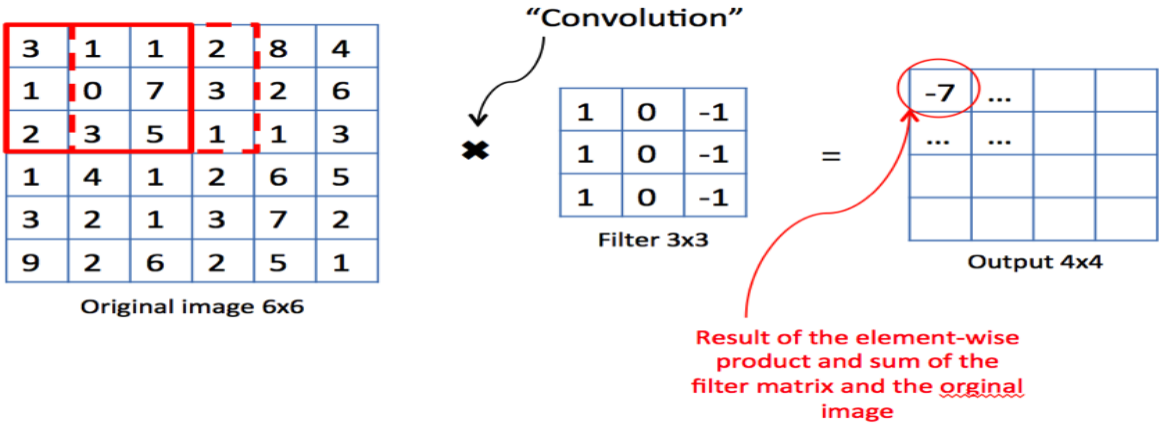


Fig 4: Process of Convolution in CNNs

If two 3 x 3 x 3 filters instead of one have been used. Then the output volume would have been 4 x 4 x 2.

Stride: Stride is the number of pixels by which we slide our filter matrix over the input matrix.

ReLU: ReLU stands for Rectified Linear Unit and is a non-linear operation. ReLU is an element wise operation (applied per pixel) and replaces all negative pixel values in the feature map by zero. The purpose of ReLU is to introduce non-linearity in ConvNet, since most of the real-world data ConvNet learn would be non-linear (Convolution is a linear operation – element wise matrix multiplication and addition, so we account for non-linearity by introducing a non-linear function like ReLU). Other non-linear functions such as tanh or sigmoid can also be used instead of ReLU, but ReLU has been found to perform better in most situations.

Pooling: Spatial Pooling (also called subsampling or downsampling) reduces the dimensionality of each feature map but retains the most important information. Spatial Pooling can be of different types: Max, Average, Sum etc.

In case of Max Pooling, a spatial neighborhood is defined (for example, a 2×2 window) and take the largest element from the rectified feature map within that window. Instead of taking the largest element one could also take the average (Average Pooling) or sum of all elements in that

window. In practice, Max Pooling has been shown to work better.

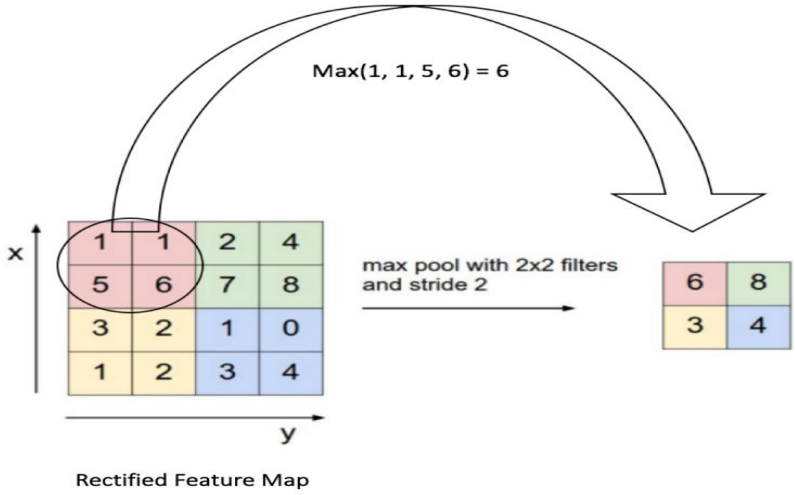


Fig 5: Process of Max Pooling in CNNs

Fully Connected Layer: The Fully Connected layer is a traditional Multi Layer Perceptron that uses a softmax activation function in the output layer (other classifiers like SVM can also be used, but will stick to softmax in this post). The term "Fully Connected" implies that every neuron in the previous layer is connected to every neuron on the next layer. The output from the convolutional and pooling layers represent high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset. The output from the convolutional and pooling layers represent high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset. The sum of output probabilities from the Fully Connected Layer is 1. This is ensured by using the Softmax as the activation function in the output layer of the Fully Connected Layer. The Softmax function takes a vector of arbitrary real-valued scores and squashes it to a vector of values between zero and one that sum to one.

6.2 Neural Architecture Search (NAS):

Developing neural network models often requires significant architecture engineering. You can sometimes get by with transfer learning, but if we want the best possible performance it's usually best to design your own network. This requires specialised skills and is challenging in general. It's a lot of trial and error and the experimentation itself is time consuming and expensive.

NAS is an algorithm that searches for the best neural network architecture. Most of the algorithms work in this following way: Start off by defining a set of "building blocks" that can possibly be used for our network. In the NAS algorithm, a controller Recurrent Neural Network (RNN) samples the building blocks, putting them together to create some kind of end-to-end architecture. This architecture generally embodies the same style as state-of-the-art networks, such as ResNets or DenseNets, but uses a much different combination and configuration of the blocks. This new network architecture is then trained to convergence to obtain some accuracy on a held-out validation set. The resulting accuracies are used to update the controller so that the controller will generate better architectures over time, perhaps by selecting better blocks or making better connections. The controller weights are updated with policy gradient.

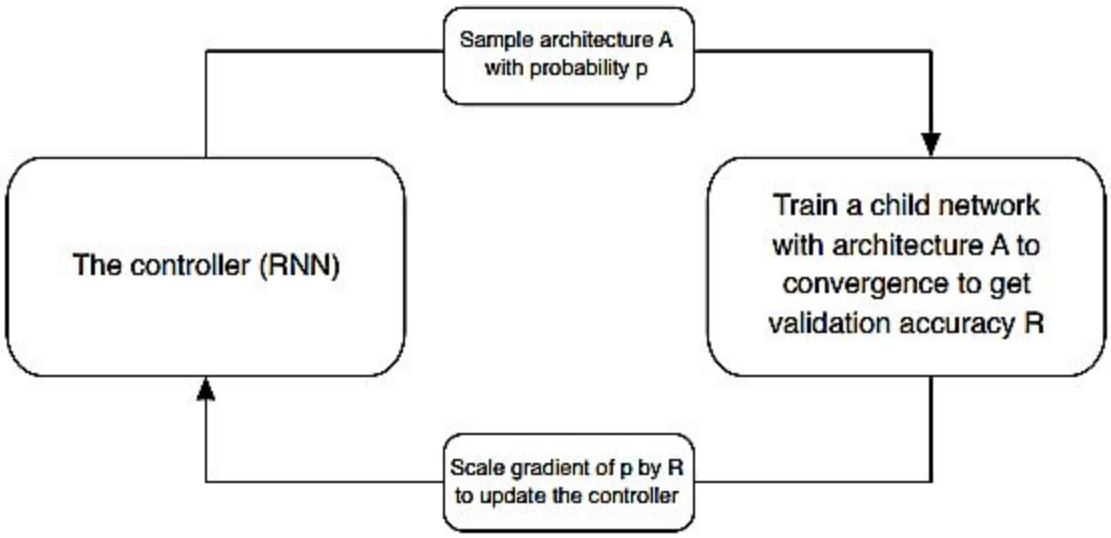


Fig 6: Working of NASnet

In simple terms: have an algorithm grab different blocks and put those blocks together to make a network. Train and test out that network. Based on our results, we can adjust the blocks we used to make the network and how you put them together.

6.3 ResNet

We found that when the deeper networks started to converge, a degradation problem got exposed: with the network depth increasing, accuracy gets saturated and then degrades rapidly.
To solve this problem instead of learning a direct mapping of x ->y with a function H(x) (A few stacked non-linear layers) a residual function was defined using F(x) = H(x) — x, which can be reframed into H(x) = F(x)+x, where F(x) and x represents the stacked non-linear layers and the identity function(input=output) respectively.The author's hypothesis is that it is easy to optimize the residual mapping function F(x) than to optimize the original, unreferenced mapping H(x).

Intuition behind Residual blocks: If the identity mapping is optimal, We can easily push the residuals to zero (F(x) = 0) than to fit an identity mapping (x, input=output) by a stack of non-linear layers. In simple language it is very easy to come up with a solution like F(x) =0 rather than F(x) = x using stack of non-linear cnn layers as function. So, this function F(x) is what the authors called Residual function.

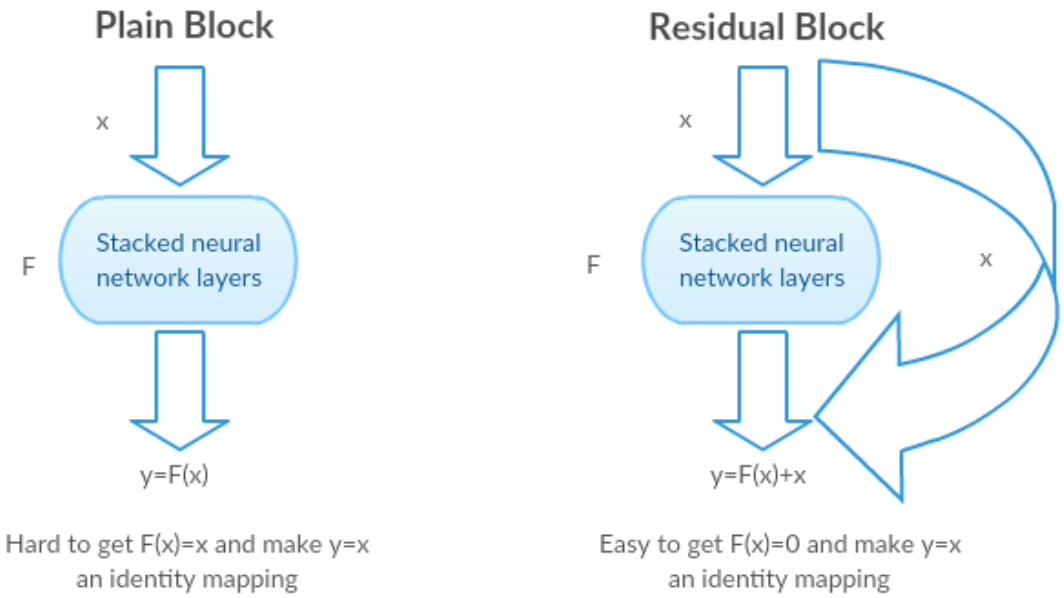


Fig 7: Visual comparison between plain and residual blocks

In deep learning networks, a residual learning framework helps to preserve good results through a network with many layers. One problem commonly cited by professionals is that with deep networks composed of many dozens of layers, accuracy can become saturated, and some degradation can occur. Some talk about a different problem called "vanishing gradient" in which the gradient fluctuations become too small to be immediately useful.

The deep residual network deals with some of these problems by using residual blocks, which take advantage of residual mapping to preserve inputs. By utilizing deep residual learning frameworks, engineers can experiment with deeper networks that have specific training challenges.

Designing the network: 3*3 filters were mostly used. Downsampling with CNN layers with stride. Global average pooling layer and a 1000-way fully-connected layer with Softmax in the end.

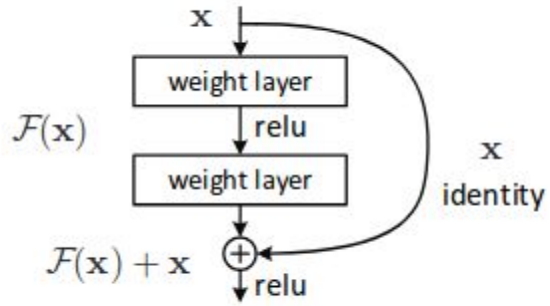


Fig 8: Residual Learning - a building block

6.4 DenseNet

The problems arise with CNNs when they go deeper. This is because the path for information from the input layer until the output layer (and for the gradient in the opposite direction) becomes so big, that they can get vanished before reaching the other side.

DenseNets simplify the connectivity pattern between layers introduced in other architectures:

- Highway Networks
- Residual Networks
- Fractal Networks

DenseNet ensures maximum information (and gradient) flow. To do this, it simply connect every layer directly with each other. Instead of drawing representational power from extremely deep or wide architectures, denseNets exploit the potential of the network through feature reuse.

Counter-intuitively, by connecting this way DenseNets require fewer parameters than an equivalent traditional CNN, as there is no need to learn redundant feature maps.

Furthermore, some variations of ResNets have proven that many layers are barely contributing and can be dropped. In fact, the number of parameters of ResNets are big because every layer has its weights to learn. Instead, DenseNets layers are very narrow (e.g. 12 filters), and they just add a small set of new feature-maps.

Another problem with very deep networks was the problems to train, because of the mentioned flow of information and gradients. DenseNets solve this issue since each layer has direct access to the gradients from the loss function and the original input image.

Structure: DenseNets do not sum the output feature maps of the layer with the incoming feature maps but concatenate them.

Consequently, the equation reshapes again into:

$$x_l = H_l (x_0 , x_1, \ldots, x_{l-1})$$

The same problem we faced on our work on ResNets, this grouping of feature maps cannot be done when the sizes of them are different. Regardless if the grouping is an addition or a concatenation. Therefore, and the same way we used for ResNets, DenseNets are divided into DenseBlocks, where the dimensions of the feature maps remains constant within a block, but the number of filters changes between them. These layers between them are called Transition Layers and take care of the downsampling applying a batch normalization, a 1x1 convolution and a 2x2 pooling layers.

Growth Rate: Since we are concatenating feature maps, this channel dimension is increasing at every layer. If we make H_l to produce k feature maps every time, then we can generalize for the l-th layer:

$$k_l = k_0 + k^* (l-1)$$

This hyperparameter k is the growth rate. The growth rate regulates how much information is added to the network each layer. We could see the feature maps as the information of the network. Every layer has access to its preceding feature maps, and therefore, to the collective knowledge. Each layer is then adding a new information to this collective knowledge, in concrete k feature maps of information.

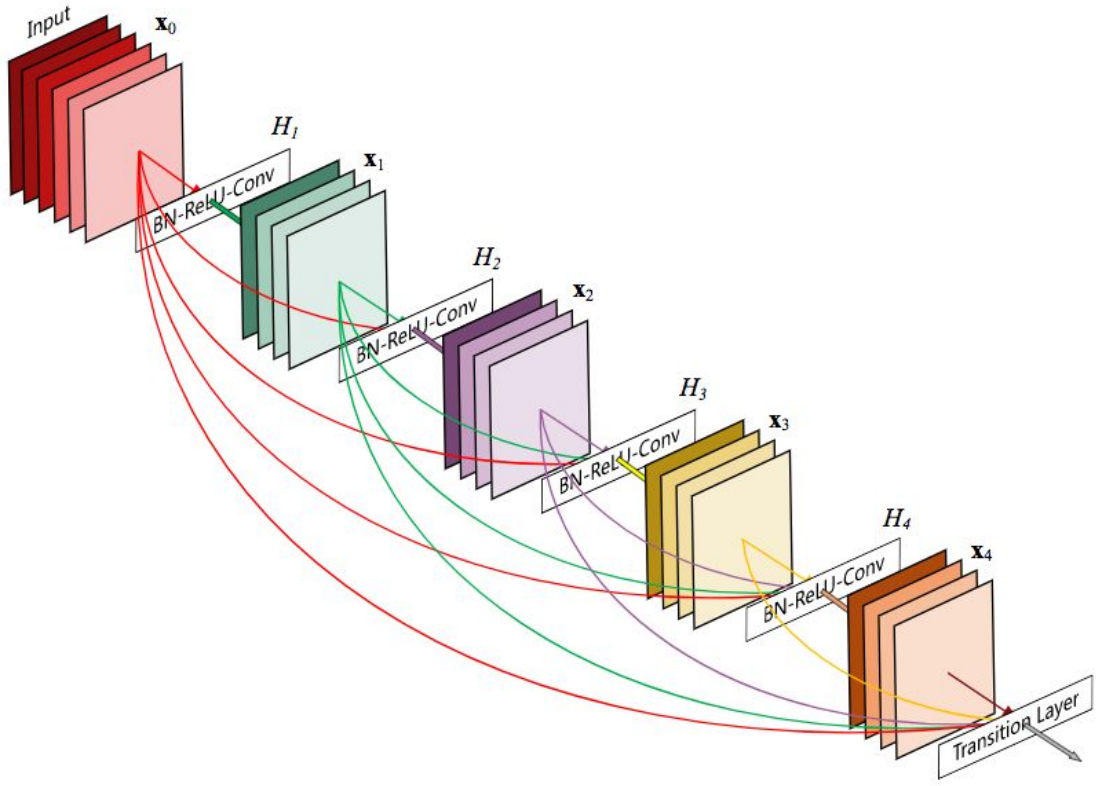


Fig 9: DenseNet Structure

6.5 MobileNet

MobileNets are based on a streamlined architecture that uses depth wise separable convolutions to build light weight deep neural networks. The MobileNet model is based on depth wise separable convolutions which is a form of factorized convolutional which factorize a standard convolution into a depthwise convolution and a 1×1 convolution called a pointwise convolution. For MobileNets the depthwise convolution applies a single filter to each input channel. The pointwise convolution then applies a 1×1 convolution to combine the outputs the depthwise convolution. A standard convolution both filters and combines inputs into a new set of outputs in one step. The depthwise separable convolution splits this into two layers, a separate layer for filtering and a separate layer for combining. This factorization has the effect of drastically reducing computation and model size.

Depthwise separable convolution are made up of two layers: depthwise convolutions and pointwise convolutions. We use depthwise convolutions to apply a single filter per each input channel (input depth). Pointwise convolution, a simple 1×1 convolution, is then used to create a linear combination of the output of the depthwise layer. MobileNets use both batchnorm and ReLU nonlinearities for both layers.

Depthwise convolution is extremely efficient relative to standard convolution. However it only filters input channels, it does not combine them to create new features. So an additional layer that computes a linear combination of the output of depthwise convolution via 1 × 1 convolution is needed in order to generate these new features. The combination of depthwise convolution and 1 × 1 (pointwise) convolution is called depthwise separable convolution.

MobileNet uses 3 × 3 depthwise separable convolutions which uses between 8 to 9 times less computation than standard convolutions at only a small reduction in accuracy.

The MobileNet structure is built on depthwise separable convolutions as mentioned in the previous section except for the first layer which is a full convolution. By defining the network in such simple terms we are able to easily explore network topologies to find a good network.

MobileNet models were trained in TensorFlow using RMSprop with asynchronous gradient descent similar to Inception V3. However, contrary to training large models we use less regularization and data augmentation techniques because small models have less trouble with overfitting. When training MobileNets we do not use side heads or label smoothing and additionally reduce the amount image of distortions by limiting the size of small crops that are used in large Inception training.

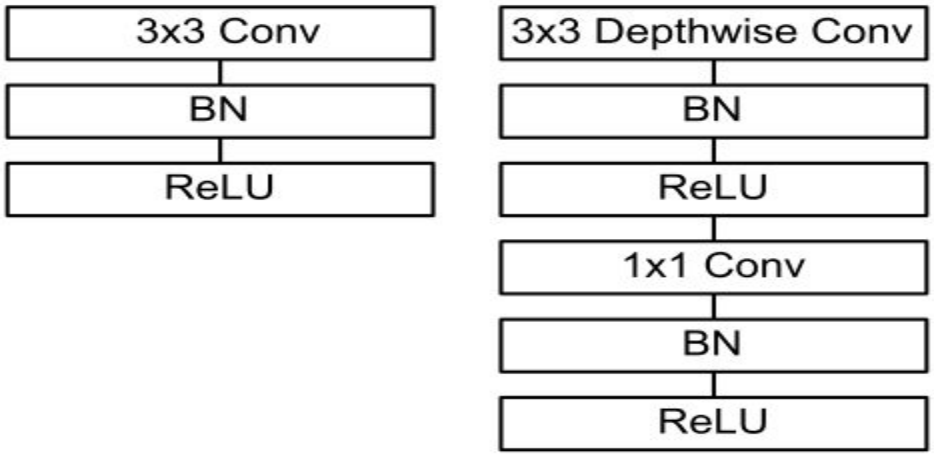


Fig 10: MobileNet Architecture showing Depthwise Convolution layer

6.6 Inception V3

The approach we are concerning in our thesis, Inception, was developed based on GoogLeNet architecture seen in ILSVRC 20143. It also took inspiration from the approach based on primate visual cortex dictated by Serre et al. which can handle multiple scales. One of the important criteria of Inception architecture is their adaption of "Network in Network" approach by Lin et al which increased the representational power of the neural networks. This had additionally saved them for computational bottlenecks by dimension reduction to 1×1 convolutions. The purpose of Inception architecture was to reduce computational resource usage in highly accurate image classification using deep learning. They had focused on finding an optimized position between the traditional way of increasing performance, which is to increase size and depth, and using sparsity in the layers based on the theoretical grounds given by Arora et al. Both the approach in their own position can cost a huge amount of computational resources. For such a deep learning system like Inception which uses fully learned filters in their 22 layer architecture, this was the main goal to achieve. They focused on the approach of Arora et al. to generate a correlation statistic analysis to generate groups of higher correlation to feed forward to the next layer. And they took the idea of multiscale analysis of visual information in their 1×1 , 3×3 and 5×5 convolution layers. All of these layers then go through dimension reduction to end up in 1×1

convolutions.

Inception V3 uses 3 techniques:

1. Factorizing Convolutions
2. Auxiliary Classifiers
3. Efficient Grid Size Reduction

1. Factorizing Convolutions

The aim of factorizing Convolutions is to reduce the number of connections/parameters without decreasing the network efficiency.

1.1. Factorization Into Smaller Convolutions

Two 3×3 convolutions replaces one 5×5 convolution as follows:

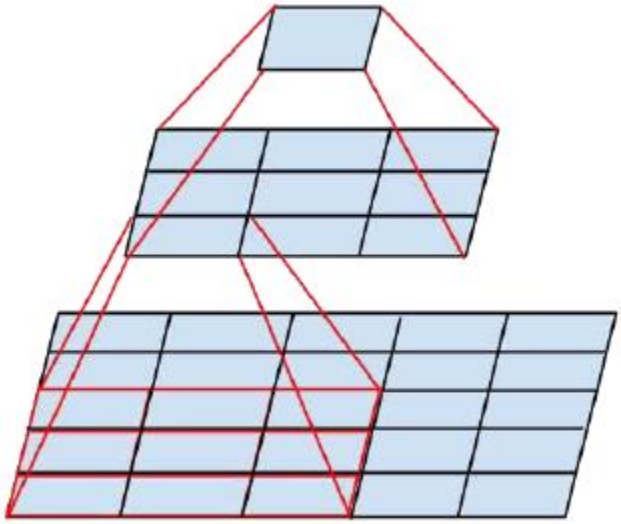


Fig 11: Two 3×3 convolutions replacing one 5×5 convolution

Two 3×3 convolutions replacing one 5×5 convolution

By using 1 layer of 5×5 filter, number of parameters = 5×5=25

By using 2 layers of 3×3 filters, number of parameters = 3×3+3×3=18

Number of parameters is reduced by 28%

With this technique, one of the new Inception modules becomes:

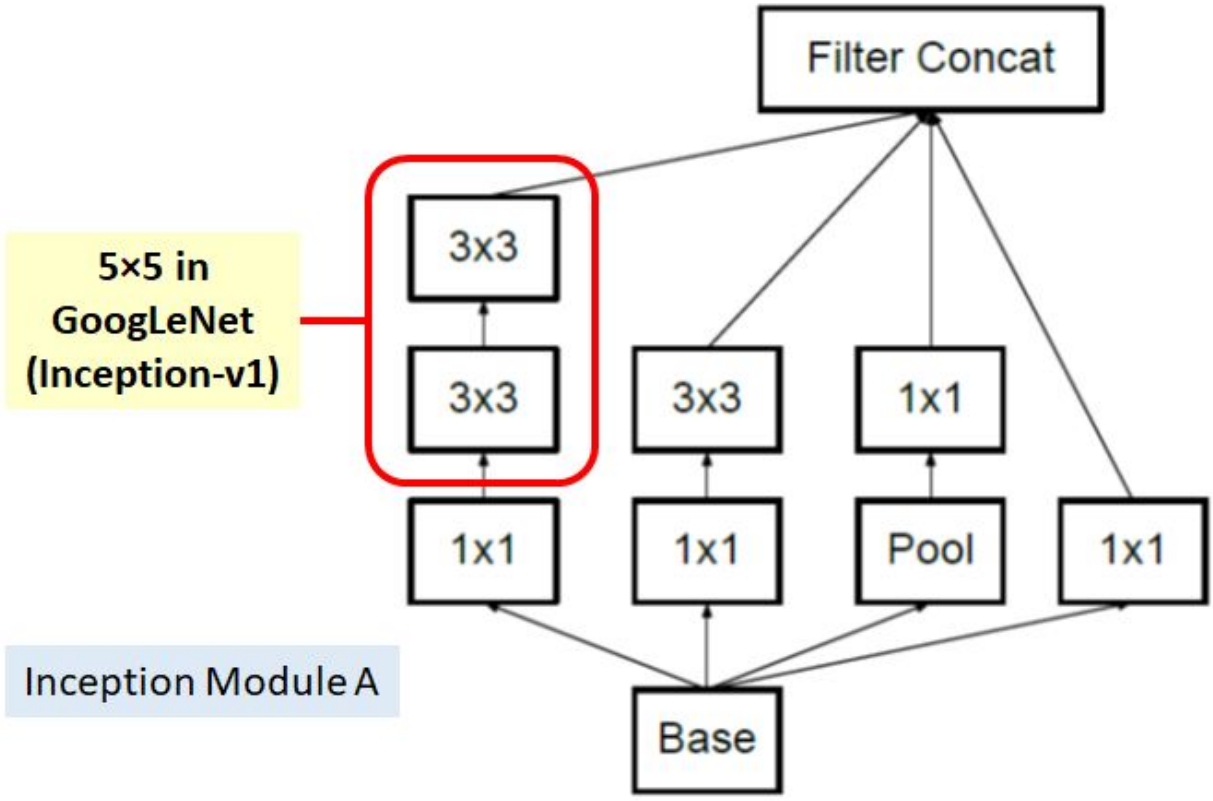


Fig 12: Inception Module A using factorization

1.2. Factorization Into Asymmetric Convolutions

One 3×1 convolution followed by one 1×3 convolution replaces one 3×3 convolution as follows:

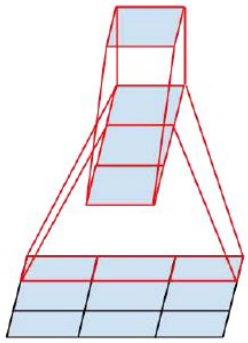


Fig 13: One 3×1 convolution followed by one 1×3 replaces one 3×3 convolution

One 3×1 convolution followed by one 1×3 convolution replaces one 3×3 convolution

By using 3×3 filter, number of parameters = 3×3=9

By using 3×1 and 1×3 filters, number of parameters = 3×1+1×3=6

Number of parameters is reduced by 33%

With this technique, one of the new Inception modules becomes:

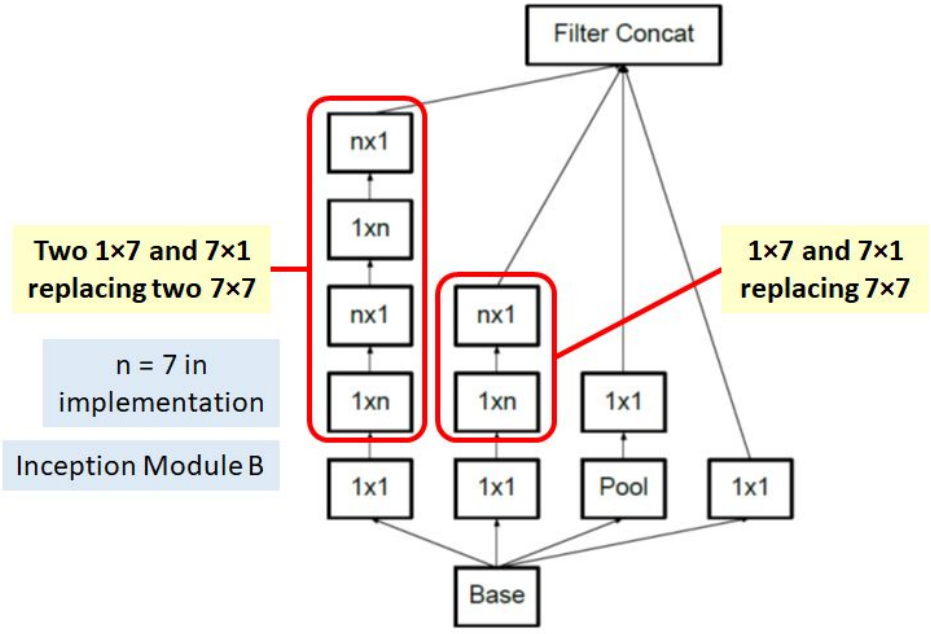


Fig 14: Inception Module B

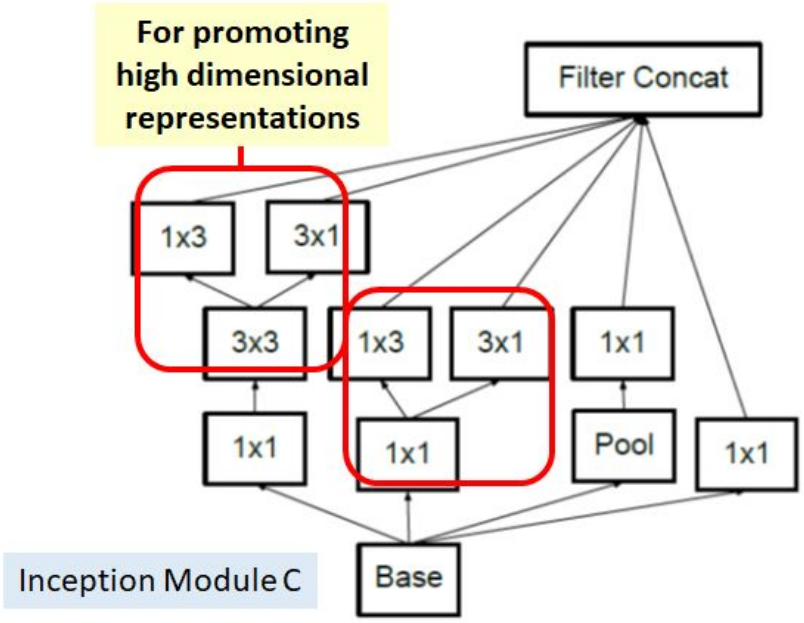


Fig 15: Inception Module C

With factorization, number of parameters is reduced for the whole network, it is less likely to be overfitting, and consequently, the network can go deeper.

2. Auxiliary Classifier

Auxiliary Classifiers were already suggested in GoogLeNet / Inception-v1. There are some modifications in Inception-v3.

Only 1 auxiliary classifier is used on the top of the last 17×17 layer, instead of using 2 auxiliary classifiers. (The overall architecture would be shown later.)

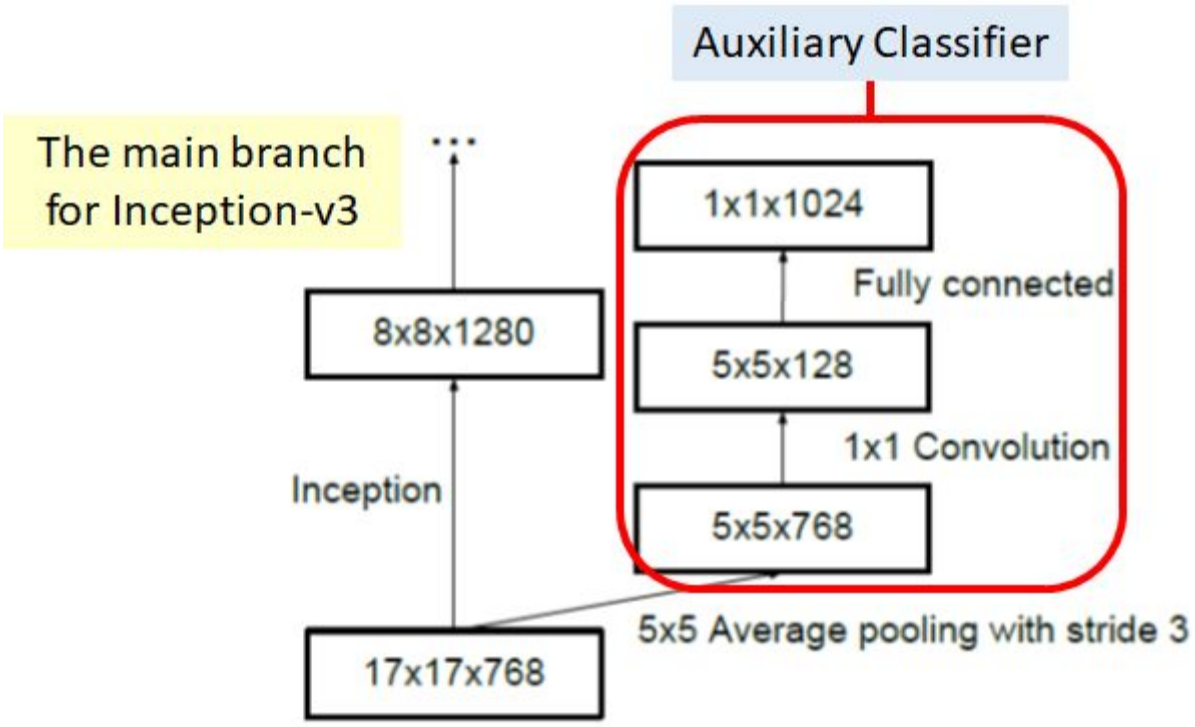


Fig 16: Main branch and auxiliary classifier of Inception-v3 architecture

Auxiliary Classifier act as a regularization. The purpose is also different. In GoogLeNet / Inception-v1, auxiliary classifiers are used for having deeper network. In Inception-v3, auxiliary classifier is used as regularizer. So, actually, in deep learning, the modules are still quite intuitive.

Batch normalization, suggested in Inception-v2, is also used in the auxiliary classifier.

3. Efficient Grid Size Reduction

Conventionally, such as AlexNet and VGGNet, the feature map downsizing is done by max pooling. But the drawback is either too greedy by max pooling followed by conv layer, or too expensive by conv layer followed by max pooling. Therefore, grid reduction is done as follows:

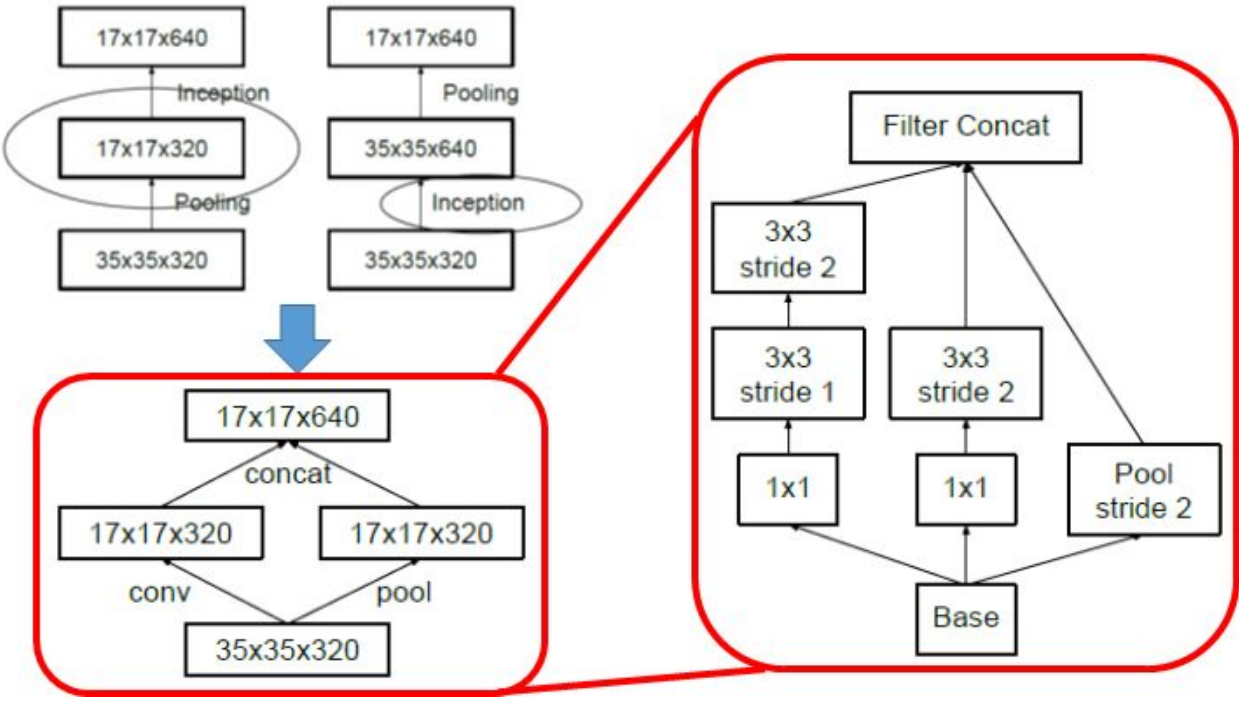


Fig 17: Conventional downsizing (Top Left), Efficient Grid Size Reduction (Bottom Left), Detailed Architecture of Efficient Grid Size Reduction (Right)

Conventional downsizing (Top Left), Efficient Grid Size Reduction (Bottom Left), Detailed Architecture of Efficient Grid Size Reduction (Right).

With the efficient grid size reduction, 320 feature maps are done by conv with stride 2. 320 feature maps are obtained by max pooling. And these 2 sets of feature maps are concatenated as 640 feature maps and go to the next level of inception module.

Less expensive and still efficient network is achieved by this efficient grid size reduction.

Inception-v3 Architecture

With 42 layers deep, the computation cost is only about 2.5 higher than that of GoogLeNet, and much more efficient than that of VGGNet.

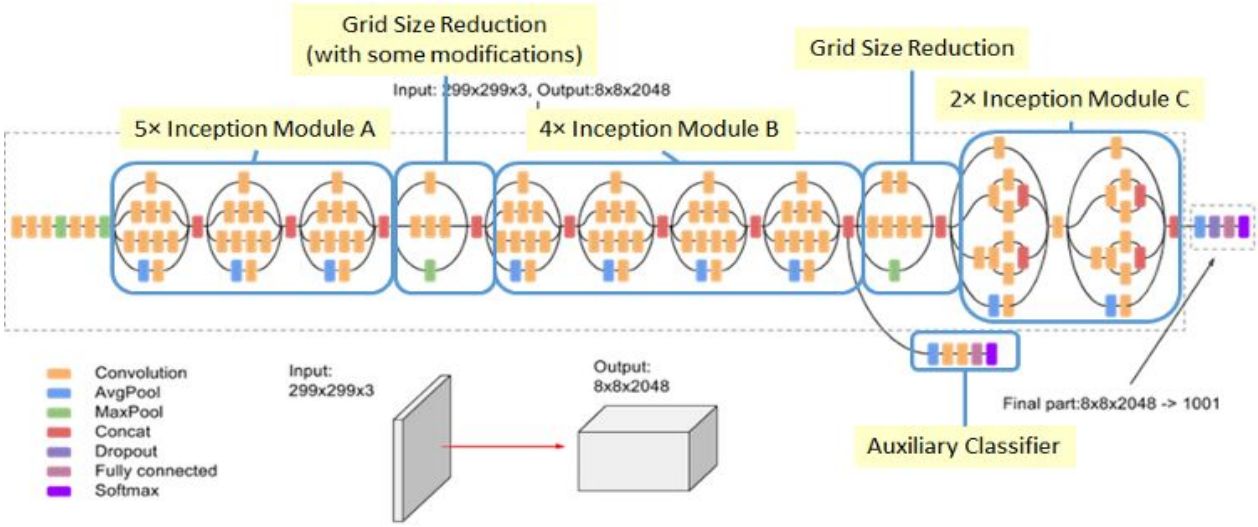


Fig 18: Inception-v3 Architecture (Batch Norm and ReLU are used after Conv)

6.7 Localisation of Abnormality using Class Activation Mapping:

The parts of the radiograph which contribute most to the model's prediction of abnormality can be visualised by using class activation mappings (CAMs).

The convolutional units of various layers of convolutional neural networks (CNNs) actually behave as object detectors despite no supervision on the location of the object was provided. Despite having this remarkable ability to localize objects in the convolutional layers, this ability is lost when fully-connected layers are used for classification. It is found that the advantages of this global average pooling layer extend beyond simply acting as a regularizer - In fact, with a little tweaking, the network can retain its remarkable localization ability until the final layer. This tweaking allows identifying easily the discriminative image regions in a single forward pass for a wide variety of tasks, even those that the network was not originally trained for.

A class activation map for a particular category indicates the discriminative image regions used by the CNN to identify that category. Just before the final output layer (softmax in the case of categorization), we perform global average pooling on the convolutional feature maps and use those as features for a fully-connected layer that produces the desired output (categorical or otherwise). We can identify the importance of the image regions by projecting back the weights

of the output layer on to the convolutional feature maps, a technique we call class activation mapping. Global average pooling outputs the spatial average of the feature map of each unit at the last convolutional layer. A weighted sum of these values is used to generate the final output. Similarly, we compute a weighted sum of the feature maps of the last convolutional layer to obtain our class activation maps.

For a given image, let $f_k(x, y)$ represent the activation of unit k in the last convolutional layer at spatial location (x, y). Then, for unit k, the result of performing global average pooling, F^k is $\Sigma_{x,y} f_k(x, y)$. Thus, for a given class c, the input to the softmax, S_c, is $\Sigma_k w_c^k F_k$ where w_c^k is the weight corresponding to class c for unit k. Essentially, w_c^k indicates the importance of F_k for class c. Finally the output of the softmax for class c, P_c is given by $(\exp(S_c)/ \Sigma_c \exp(Sc))$. Here we ignore the bias term: we explicitly set the input bias of the softmax to 0 as it has little to no impact on the classification performance.

By plugging $F_k = \Sigma_{x,y} f_k(x, y)$ into the class score, S_c, we obtain

$$S_c = \Sigma_k w_c^k \Sigma_{x,y} f_k(x, y)$$
$$= \Sigma_{x,y} \Sigma_k w_c^k f_k(x, y)$$

We define M_c as the class activation map for class c, where each spatial element is given by

$$M_c(x, y) = \Sigma_k w_c^k f_k(x, y)$$

Thus, $S_c = \Sigma_{x,y} M_c(x, y)$, and hence $M_c(x, y)$ directly indicates the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c. f_k is the map of the presence of visual pattern. The class activation map is simply a weighted linear sum of the presence of these visual patterns at different spatial locations. By simply upsampling the class activation map to the size of the input image, we can identify the image regions most relevant to the particular category.

In order to perform localization, we need to generate a bounding box and its associated object category. To generate a bounding box from the CAMs, we use a simple thresholding technique to segment the heatmap. We first segment the regions of which the value is above 20% of the max value of the CAM. Then we take the bounding box that covers the largest connected component in the segmentation map.

CHAPTER 7
RESULTS

Results of various models on the dataset:

The following table shows the comparison of results of different DL models on the dataset in tabular form:

Body Parts	ResNet-18	MobileNet	DenseNet	InceptionV3	NASNet
Wrist	75%	62%	77%	81%	87%
Shoulder	69%	56%	71%	73%	72%
Forearm	73%	57%	72%	75%	71%
Elbow	78%	64%	79%	82%	81%
Finger	71%	55%	71%	74%	75%
Humerus	78%	64%	81%	82%	80%
Hand	68%	57%	71%	69%	67%

Table 2 : Comparison of results between different DL models

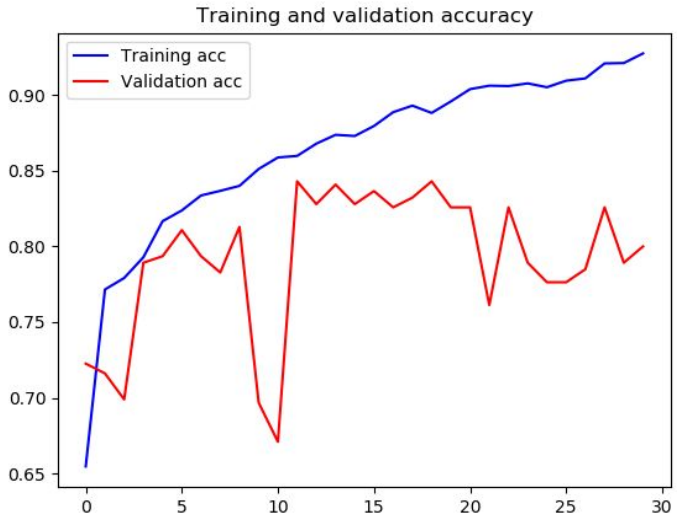


Fig 19: Inception-v3 training and validation accuracy for elbow part

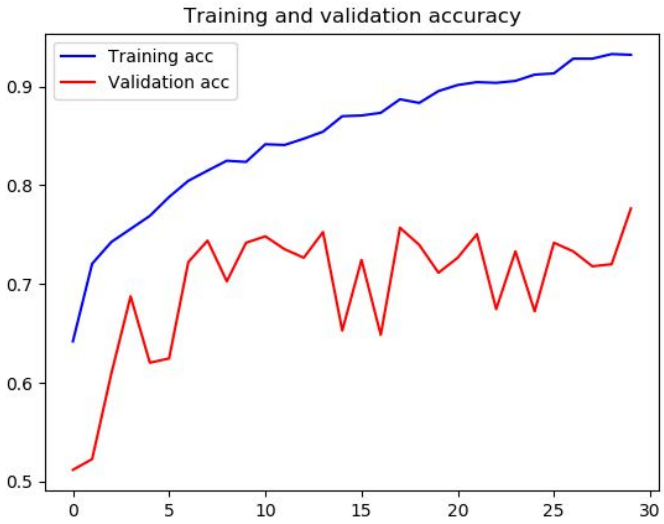


Fig 20: Inception-v3 training and validation accuracy for finger part

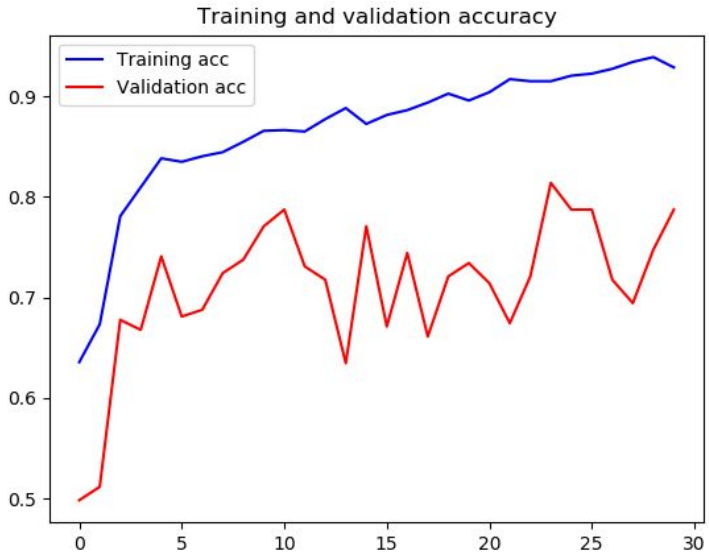


Fig 21: Inception-v3 training and validation accuracy for forearm part

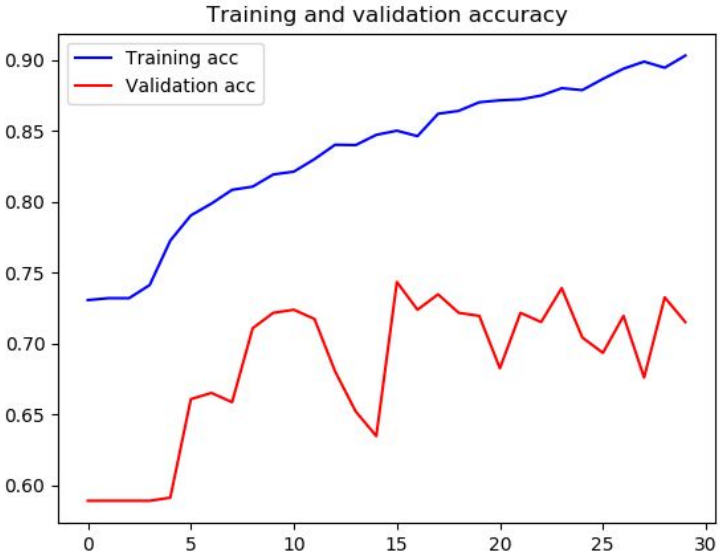


Fig 22: Inception-v3 training and validation accuracy for hand part

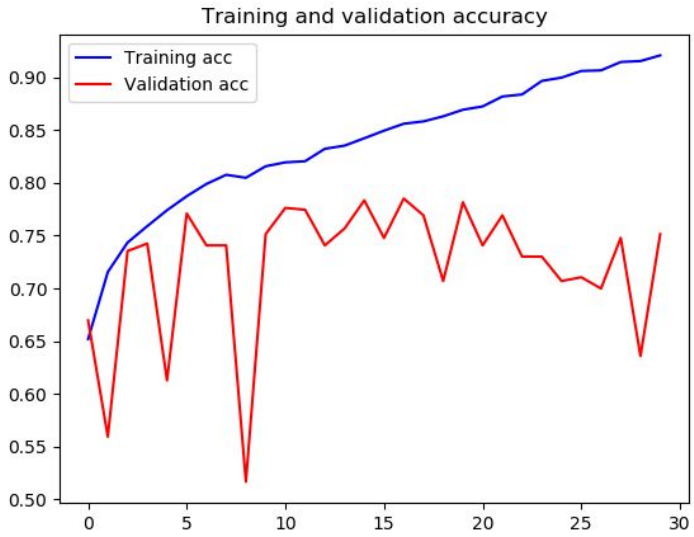


Fig 23: Inception-v3 training and validation accuracy for shoulder part

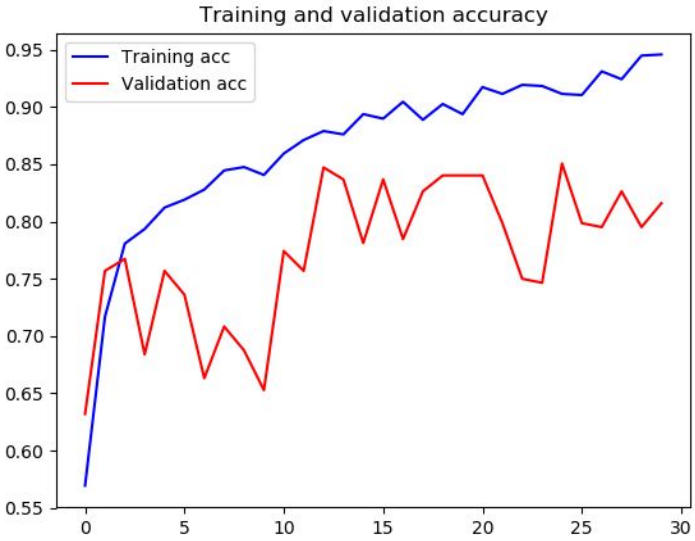


Fig 24: Inception-v3 training and validation accuracy for humerus part

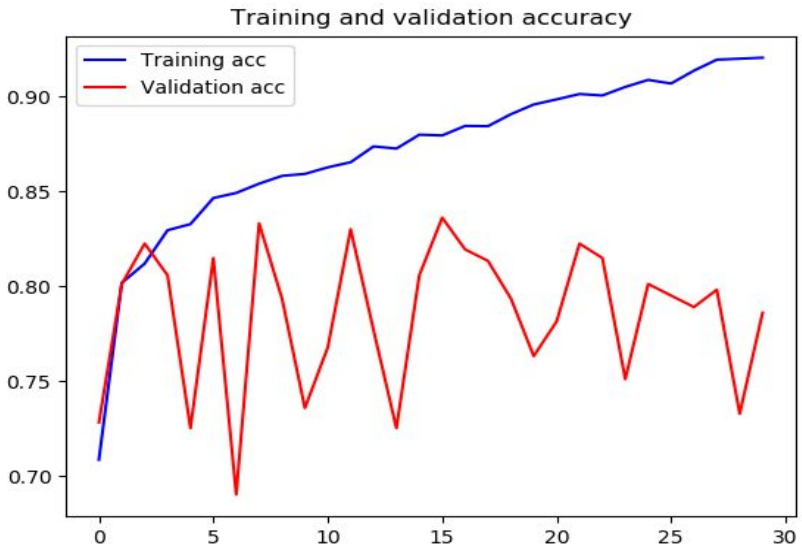


Fig 25: Inception-v3 training and validation accuracy for wrist part

Localisation of abnormality:

Using Class Activation Mapping (CAM), we have successfully localised the abnormality which is helpful for localisation. which can be seen below:

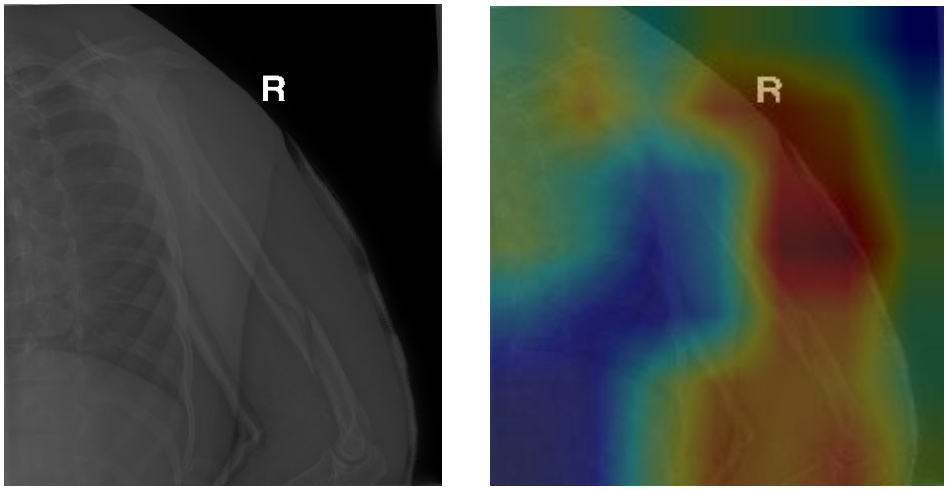


Fig 26: Localisation of abnormality (right) in the original image (left)

CHAPTER-8

CONCLUSION AND FINDINGS

In this work we have used different algorithms for classification of the X-ray images into normal and abnormal, and by far we got the best results using InceptionV3. Firstly, we used CNN but we saw that as the number of epochs were increasing the accuracy started getting saturated and afterwards degraded rapidly so we needed a shallow network for this and residual net fitted completely this criteria as it adds a residual part of the previous layer instead of direct mapping.

After training ResNet we found that the number of parameters in ResNet was very high hence it needed a large time to train many of these layers were barely contributing hence there was introduction to DenseNet whose layers are very narrow and add a small set of new feature map. Further, we tried Inception V3 whose parameter size was even reduced due to the factorizing convolutions and accuracy was even better. The last method we tried is NasNet which improves uses reinforcement learning to get an optimized model but as we had limited resources the number of layers taken were maximum 10 hence the results obtained NasNet can still be better.

REFERENCES

[1]AIMI. Artificial intelligence in medicine & imaging: Available labeled medical datasets. https: //aimi.stanford.edu/available-labeled-medical-datasets. [Online; accessed 2-December-2017].

[2]Berlin, Leonard. Liability of interpreting too many radiographs. American Journal of Roentgenology, 175(1):17–22, 2000.

[3]Bhargavan, Mythreyi and Sunshine, Jonathan H. Utilization of radiology services in the united states: levels and trends in modalities, regions, and populations. Radiology, 234(3):824–832, 2005.

[4]Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 248–255. IEEE, 2009.

[5]Gertych, Arkadiusz, Zhang, Aifeng, Sayre, James, Pospiech-Kurkowska, Sylwia, and Huang, HK. Bone age assessment of children using a digital hand atlas. Computerized Medical Imaging and Graphics, 31(4):322–331, 2007.

[6]Huang, Gao, Liu, Zhuang, Weinberger, Kilian Q, and van der Maaten, Laurens. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.

[7]Jaeger, Stefan, Candemir, Sema, Antani, Sameer, Wang, Yi-Xiang J, Lu, Pu-Xuan, and Thoma, George. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery, 4(6):475, 2014.

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

[9]Lu, Ying, Zhao, Shoujun, Chu, Philip W, and Arenson, Ronald L. An update survey of academic radiologists' clinical productivity. Journal of the American College of Radiology, 5(7):817–826, 2008.

[10]McHugh, Mary L. Interrater reliability: the kappa statistic. Biochemia Medica, 22(3):276–282,October2012. ISSN 1330-0962. URL https://www.ncbi.nlm.nih.gov/pmc/articles/

PMC3900052/. Nakajima, Yasuo, Yamada, Kei, Imamura, Keiko, and Kobayashi, Kazuko. Radiologist supply and workload: international comparison. Radiation Medicine, 26(8):455–465, 2008.

[11]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, 2017b.

[12]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. arXiv preprint arXiv:1705.02315, 2017.

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