Convolutional Neural Network Based Thermal Image Classification

Author

Qirat Ashfaq

00000274771

Supervisor

Dr. Muhammad Usman Akram

DEPARTMENT OF COMPUTER SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

MAY, 2021

Convolutional Neural Network Based Thermal Image Classification

Author

Qirat Ashfaq

00000274771

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Software Engineering

Thesis Supervisor

Dr. Muhammad Usman Akram

Thesis Supervisor's Signature: _____

DEPARTMENT OF COMPUTER SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

MAY, 2021

Declaration

I certify that this research work titled "*Convolutional Neural Network Based Thermal Image Classification"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student
Qirat Ashfaq
00000274771

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
Qirat Ashfaq
00000274771

Signature of Supervisor
Dr.Muhammad Usman Akram

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

Acknowledgements

All praise and glory to Almighty Allah (the most glorified, the highest) who gave me the courage, patience, knowledge and ability to carry out this work and to persevere and complete it satisfactorily. Undoubtedly, HE eased my way and without HIS blessings I can achieve nothing.

I would like to express my sincere gratitude to my advisor Dr. Muhammad Usman Akram for boosting my morale and for his continual assistance, motivation, dedication and invaluable guidance in my quest for knowledge. I am blessed to have such a co-operative advisor and kind mentor for my research.

Along with my advisor, I would like to acknowledge my entire thesis committee: Dr. Sajid Gul Khawaja, and Dr. Arslan Shaukat for their cooperation and prudent suggestions.

My acknowledgement would be incomplete without thanking the biggest source of my strength, my family. I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

Finally, I would like to express my gratitude to all my friends and the individuals who have encouraged and supported me through this entire period.

Dedicated to my exceptional parents: ***Ashfaq Ahmed Tahir (Late) & Samina Ashfaq****, and adored husband (Waqas Ahmed) whose tremendous support and cooperation led me to this accomplishment*

Abstract

Classification of Thermal Images has been extensively used for its significant applications in many fields. There are many problems with the visible spectrum like object shadows, clothes or the body of human being matches the background and different lighting conditions. These limitations are overcome by using thermal imaging. Each and every object emits heat (Infrared energy) according to its temperature. Normally the hotter object emits more radiation then the colder one. As all objects have mostly different temperature so thermal camera detects them and these objects get appear as distinct objects. In the start thermal imaging was used by military for detection, recognition and identification of enemy personnel and equipment. Now a days it is extensively used in detection of face, self-driving car, detection of pedestrian and it also have application in the field of environmental work that is monitoring for energy conservation and pollution control. This research presents a novel study for the classification of thermal images using convolutional neural networks (CNN). Research focused on developing a framework that detects multiple thermal objects using CNN. Developed a framework based on deep learning Inception v3 model; work with thermal images that are captured by Seek Thermal and FLIR. For training and testing of the model two datasets are used that include three classes' cat, car, and man. For FLIR dataset the highest accuracy achieved is 98.91% and for Seek thermal dataset highest accuracy achieved is 100%. A comparison of proposed framework with some other CNN models (DenseNet, MobileNet and YOLOv4), with customized CNN model and with a conventional model is also presented. The results of proposed framework and comparison with other models prove that proposed framework is effective for the classification of thermal images.

Key Words: *Infrared energy, convolutional neural networks, deep learning, DenseNet, MobileNet, YOLOv4*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

Classification of thermal images has been extensively used for its significant applications in many fields. There are many problems with the visible spectrum like object shadows, clothes or the body of human being matches the background and different lighting conditions. These limitations are overcome by using thermal imaging. Each and every object emits heat (Infrared energy) according to its temperature. Normally the hotter object emits more radiation then the colder one [3]. As all objects have mostly different temperature so thermal camera detects them and these objects get appear as distinct objects. In the start thermal imaging was used by military for detection, recognition and identification of enemy personnel and equipment. Now a days it is extensively used in detection of face, self-driving car, detection of pedestrian and it also have application in the field of environmental work that is monitoring for energy conservation and pollution control. This research paper presents a novel study for the classification of thermal images using CNN. Our research focused on developing a method that detects multiple thermal objects using CNN. We have developed a model based on deep learning inception v3; work with thermal images that are captured by Seek Thermal and FLIR. For the training and testing of the models two datasets are used that include three classes' cat, car, and human.

1.1 Motivation

Currently, most of the study is in visible imaging that has a lot of limitations. Some conventional methods did thermal image classification but no work is done on the usage of deep learning for the classification of thermal images. Models of the deep learning achieve the highest accuracy; sometimes surpass the performance of humans [28]. Convolutional neural network is an algorithm of deep learning that is used for image classification. In CNN the requirement of pre-processing is much less than the other classification algorithms [6]. So, by keeping in view the accuracy of deep learning in the domain of classification of images and benefits of thermal image classification, we conducted a research on developing a model based on deep learning; work with thermal images that are captured by Seek Thermal and FLIR. We will use our own dataset no online dataset related to thermal objects man, car and cat is found.

1.2 Problem Statement

In view of the advance security accurate classification of thermal images have a lot of significance. The purpose of this research is to collect thermal images dataset and apply convolutional neural network for object classification.

1.3 Aims and Objectives

Major objectives of the research are as follow:

- To provide a good dataset of thermal images for object classification to the research community
- To review the recent work in the classification of thermal images.
- To classify the thermal images of multiple objects.
- To develop a model based on deep learning that work with thermal images that are captured by Seek Thermal and FLIR with high accuracy

1.4 Major research groups/ Universities working in this area

- University of Glasgow
- University of Silesia
- Plymounth University
- Brock University
- University of Sussex
- Lund University

1.5 Structure of Thesis

This work is structured as follows:

Chapter 2 covers the importance of thermal imaging and its applications

Chapter 3 gives the review of the significant work and the literature done by researchers in past few years for the thermal image classification.

Chapter 4 presents the proposed methodology in detail.

Chapter 5 gives the introduction of the databases used for the evaluation purposes. Experimental results are discussed in details with all the required tables and figures.

Chapter 6 discloses the future scope of the research and concludes the thesis

CHAPTER 2: Thermal Imaging and its Application

2.1 Thermal Imaging/FLIR/IR Images:

Thermal imaging is the procedure of converting infrared energy (heat) into the visible images for the analysis of the surroundings. In an image captured by the thermal camera it illustrates the dispersal of temperature differences.

i. Brief explanation on the working of thermal imaging:

Each and every object emits heat (Infrared energy) according to its temperature. The infrared energy that is emitted by the object is called its heat signature. Normally the hotter object emits more radiation then the colder one [3].

Thermal camera also known as thermal imager is a heat sensor that has the capability of detecting the slightest difference in the temperature. This camera gathers all the infrared radiation from the objects that are there in the scene and on the basis of the temperature deference information it forms the electronic image. As all objects have mostly different temperature so thermal camera detects them and these objects get appear as distinct objects. There are two types of thermal images: mostly grayscale in which cold objects are represented by using block color and hot objects with white color and the variations between these two are represented by the depth of the grey [5]. However some thermal cameras add colors to the images for the better identification of objects.

An infrared camera produces IR images it is a non-contact device it detects the infrared energy (heat) and then converts it into an electronic signal and then this signal is processed for the production of thermal images or IR images. Heat that is sensed by the IR camera can be measured or quantified very precisely.

ii. EM Band in which Thermal Images are acquired:

The electromagnetic (EM) band is a range of EM radiations of all types. Radiation is an energy that move and outspreads like light coming from the lamp is the visible light and waves coming out from the radio station are the radio waves. There are different type of radiation that combine form EM band like infrared rays, microwaves, gamma rays, X-rays and visible spectrum. [45]Wavelength, frequency and energy are the terms use to express

electromagnetic radiations. Hertz or cycles per seconds are the unit of frequency. A meter is the unit of wavelength. Electron volt is the unit of Energy. Each of three quantities frequency, energy and wavelength are related to each other for describing the EM radiation in the precise way. The far spectral infrared range for the thermal images are 3 – 5 µm or 8 – 12 µm. [34] Spectral infrared range: 0.7 - 300 µm wavelength. Following are the bands in which this region is divided:

- Near Infrared (NIR) range is 0.7 - 1.5 µm.
- Short Wavelength Infrared (SWIR) range is 1.5 to 3 µm.
- Mid Wavelength Infrared (MWIR) range is 3 to 8 µm.
- Long Wavelength Infrared (LWIR) range is 8 to 15 µm.
- Far Infrared (FIR) range is more than 15 µm.

The SWIR and NIR are main IR component of solar radiation that is reflected from earth surface. The LWIR and MWIR are the Thermal IR. [35]

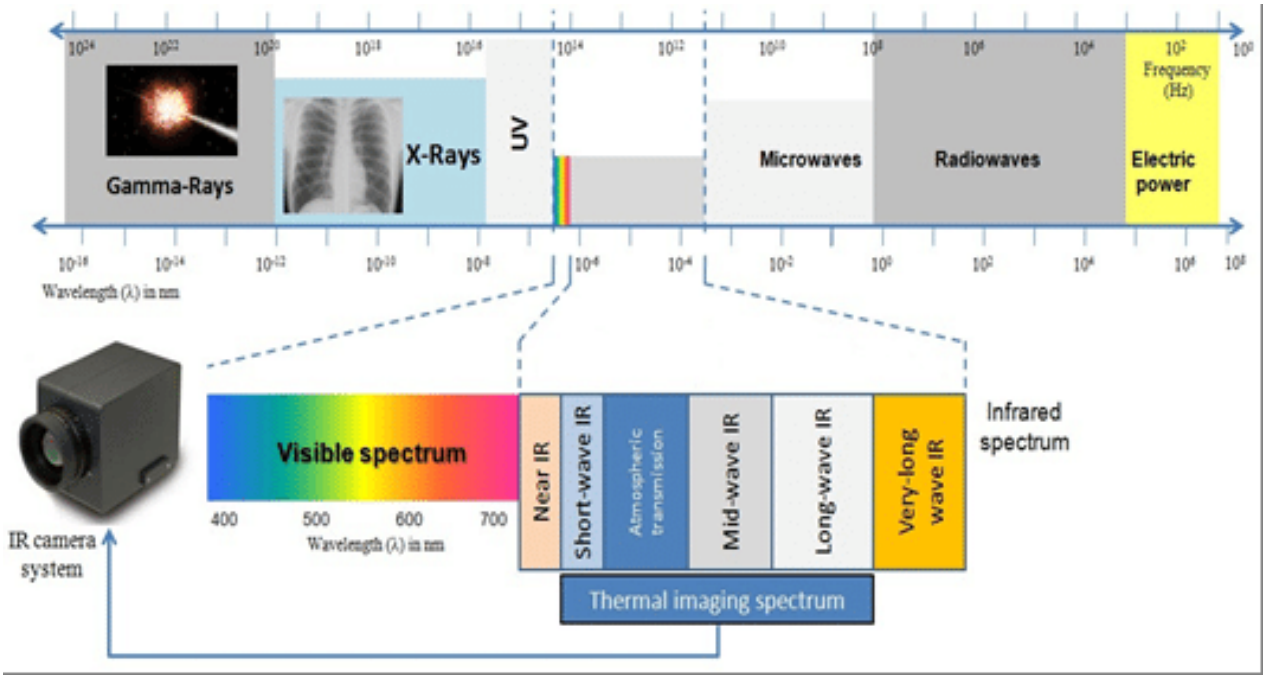


Figure 2.1: EM showing IR spectrum [36]

2.2 Thermal cameras

A thermal camera is a device that senses the heat (infrared energy) and then converts it into the visible image. Let's see the details of what the science of thermal cameras and how it enables us to see the heat that is invisible world for us. First of all working of the regular cameras is different from the thermal cameras. The basic principle of human eye and regular

camera is same, that is when something is hit by the visible light then it bounces back , this reflected light enter into the detector and turns into an image.

Thermal imagers or cameras make an image not from the visible light but from the heat signature of the object. Infrared or thermal energy is also known as heat. Light and the heat both are the EM spectrum's part but some cameras detect visible light and some detects heat. Infrared energy is captured by the thermal cameras and then data obtained from this is used to create a visible image by using analog or digital video outputs.

Lens, processing electronics, a thermal sensor and mechanical housing is used to make a thermal camera. The lenses of the thermal camera focus the infrared energy onto the sensor. The sensors are of different pixel configurations from 80 x 60 to 1280 x 1024 pixel or more than that. This pixel configuration is the resolution of the camera.[46]

Thermal cameras have to detect energy that has much larger wavelength in comparison with the visible light due to which the resolutions of thermal cameras are low as compare to visible light cameras. As a result, a visible camera of same mechanical size has much higher resolution then the thermal cameras. For choosing the thermal cameras some of the important specifications are field of view, spectral range, resolution, thermal sensitivity and focus.

i. Major companies provide thermal cameras

It is estimated that the global market of the **thermal imaging** will rise from USD 3.4 billion in 2020 to the 4.6 billion USD till 2025, at a 6.2% of CAGR. The investments of R and D by the government, capital firms and companies is the key factor of fueling the thermal imaging market's growth, these investments are for developing novel thermal imaging solutions and to increase in the implementation of thermal imaging in the industry of automotive [43]. Some of the major companies that provide thermal imaging cameras are listed below: [37] [38]

- FLIR Systems – One of the leading company that manufacture thermal images infrared cameras.
- FLUKE
- Seek Thermal

- Trijicon
- HT Instruments
- Perfect Prime

ii. Best Thermal Imaging Camera Models:

Why there is a need for best thermal imaging camera? There is need of best thermal imaging camera because it allows seeing and making the accurate measurement of temperature difference from the safest distance. For the identification of object in the dark or the obscured places the best thermal cameras are very useful like used to rescue a child who is lost in a dark or also help in controlling your heating bills. It will also help in Covid19 by identifying person with high feverish temperature. [39][40]

- FLIR TG165 Spot
- FLIR E4: Compact
- FLIR C2 Compact
- Seek Thermal iOS-Apple
- Fluke TIS20 9HZ
- FLIR E8: Compact 320 x 240
- FLIR C3 Pocket
- Seek Thermal Reveal
- Perfect-Prime IR0002
- FLIR TG56 Spot
- Seek Thermal XR Imager
- FLIR ONE Pro Thermal Imagers

Thermal imaging cameras that are used for the collection of dataset for the presented research are FLIR and Seek Thermal.

i. FLIR - Forward Looking Infrared Radar:

FLIR Systems is the largest commercial company of the world that is specialized in the production and the design of imaging sensors, thermal imaging cameras and components. FLIR use a thermo graphic camera to sense the infrared radiation. [44]

ii. Seek Thermal:

Low-cost and high-resolution thermal imaging cameras are manufactured and engineered by Seek Thermal. A pioneer founded this industry that spent 40 years in advancing the state of

professional grade and military thermal technologies. Seek Thermal has developed products at the competitive prices so that it is accessible to the more end users, the products of the company serve the law enforcement, firefighting and commercial markets. [41]

2.3 Applications of thermal Imaging:

The impending applications of thermal imaging are nearly limitless. In the start thermal imaging was used for the military for detection, recognition and identification of enemy personnel and equipment, now widely used .Some of the applications are listed below: [33]

- Now days it is extensively used in border patrol crossings and fire safety activities.
- It also have application in the field of environmental work that is monitoring for energy conservation and pollution control.
- It has different construction, medical and industrial applications.
- Police force makes use of it for the management of surveillance activities, locate and capture suspects, for the investigation of the crime scenes and for the rescue operations.
- To identify overheating parts and joints power line technicians use it.
- Building construction technicians use it for the identification of faulty thermal insulation.
- Thermo graphic imaging used for the monitoring of physiological activities like fever.
- Infrared light have greater penetration power then visible light in smog and fog.
- In a chemical reaction or in a piece of machinery excess cold or heat can be detected by using it.
- It will also help in Covid19 by identifying person with high feverish temperature

CHAPTER 3: LITERATURE REVIEW

Classification of the images has been extensively studied now a days but most of the study is in the domain of visible imaging, that requires adequate light environment. But in real scenario there is a need of image classification in night environment, when street lights are out of power or in outdoor environment where there is no light. These limitations are overcome by using thermal imaging. Some conventional methods did thermal image classification but no work is done on the usage of deep learning for the thermal image classification. Some international literature was found related to thermal image classification but no literature is related to usage of Model developed by using inception v3 for the classification of thermal images. So no resources about documentation were available for the comparison. This chapter will summarize all those valuable researches in this domain.

3.1 Conventional Methods for image classification:

Paramount problems with visible spectrum is the object shadows, object matches the background and dependency on the condition of light. Thermal IR cameras overcome these limitations of visible spectrum by gathering all the infrared radiation (Heat) from the objects that are there in the view and on the basis of the temperature deference information it forms the electronic image. J. E. Jackson and L. Chermak [4] proposed a first utilization of thermal signatures that are regularly occurring and that was produce by a moving platform, the paradigm was defined as PUGTIFs (passively user-generated thermal invariant features). Framework consisting of three methods will be suitable for PUGTIFs, methods were: To solve monocular localization footprint size that is known is used, second to determine heading orientation the consistent spatial pattern uses and third in ATFD (automated thermal footprint detector) all these principles were combined for achieving feature detection or segmentation. PUGTIFs detection was evaluated in the lab for the four environments (grass, sand, carpet and grass). After those typical methods of image segmentation was compared to ATFD and found that ATFD is better than others. M. Haider et al. * [10] had taken electrical equipment's thermal images and then converted it to model of HSV for further processing. Most of the techniques used grey scale images but in this study hue region had been considered and a comparative analysis was presented for different methods of automatic thresholding that include Otsu, Roberts and Prewitt. This comparison was done for defected electrical equipment's thermal images like solar panel. Total samples of the thermal images

of the different electrical equipment were about 40. Standard grey scale approach have more classification accuracy and have better feature extraction competence but proposed technique gave better results for all images as compared to it. Feature detection is a very important part of the applications of computer vision, however environmental changes, light and distance greatly affect the quality of features and make its identification difficult. Hossen et al. * [21] Investigated the recognition and tracking of activities in the thermal infrared videos and made the improvement in pose segmentation by using universal segmentation technique and feature extraction technique. For the collection of non-repetitive and repetitive activities GEI- Gait Energy Images had been developed. From the sequences of GEI of each activity seven invariant moments' features were extracted and then concatenated with feature vector. For feature vectors classification Naïve Bayesians classifier was used. Data of real thermal image of repetitive human activity was recorded using FLIR A40 long-wave infrared. An Accuracy of 76% was achieved. D. Kim et al. *[20] worked on the simple and the fast approach for the pedestrian tracking and detection on the basis of shape features. A temperature-based threshold method was used to detect the pedestrian regions in an image. For the estimation of precise shape and location the background region elimination was done and then for the extraction of shape feature it was fed into the tracking step. HOG was extracted as feature of local shape. For the calculation of transition score between the current and previous frame's bounded boxes feature distance was measured, after that bounding boxes that were closest were linked together. The smaller the distance between feature lead to higher transition scores. Dataset OTCBVS was used. Performed multi pedestrian tracking and detection effectively. C. K. Kyal et al. *[12] presented an effective method for the detection of human face in thermal imaging. Histogram plot was used for the feature extraction and face detection; some other techniques were also used like Morphological operation, thresholding, and object boundary analysis. For enhancement of algorithm performance, the reduction of computation time and for achieving parallelism Message passing Interface model was used. Dataset of more than 500 images were obtained in real time using FLIR 60101-0301 Model i7. In a parallel environment algorithm was much faster as compared to sequential environment processing time was 0.11 units? Their future work contains proposed algorithm's expansion using deep learning and machine learning techniques. For the detection of new physiological and anatomical face information it also includes the use of IR thermal sensors. G. Lu et al. *[11] proposed an efficient classification algorithm that combined multiple features of image for enhancing the accuracy of classification. Feature

fusion technique was used in which four features were extracted by using different techniques for enhancing the image classification that were Edge orientation, Hu moment, Temperature feature, and Bag-of-words based on corner feature. To maximize multiple features' fusion performance, optimization of the objective function needs to be done. By the objective function's optimization the large amount of thermal image information will be utilized for the classification of images into correct group. Xenics Gobi 640 GigE infrared camera uncooled was used for collecting the 3900 images. 90.1% accuracy was achieved. Jian-Feng Shi et al. * [18] evaluated the SURF, AKAZE, SIFT, ORB, and BRIEF image features. On the various image features multiple tests were conducted. First data set that was used for the test was the Oxford image of Graffiti and Boat that was introduced by Schmid and Mikolajczyk. The second test was performed on the images that were captured by a ICI-9320P – thermal camera. Executable build from OpenCV 3.0.0 was used for the computation of results. For noise variations, viewpoint and rotational the yBRIEF descriptor showed the much higher performance than the other descriptors of BRIEF family. Due to which it was concluded that to gain greater precision binary pattern adjustments need to be used. HOG was used along with SVM to identify different targets, by merge other image processing methods this method has been authenticate for identification of objects. D. Zhou et al. *[19] presented a research to determine the presence of deer in the thermal images. HOG – Histogram of oriented gradient and SVM- support vector machine method was used for identification of deer. SVM was used for generating deer description called descriptor and then descriptor was compared with HOG of image. The result obtained from comparison told about the presence of deer in current image. For the improvement of accuracy, the second training was performed, where for training samples false positive images were used. Images of the deer were obtained in the Lake Superior Zoo in Duluth, MN USA. The field and lab test results had showed 85% accuracy. Visible images of object provide appearance patterns and geometric while sensitive to light while thermal images are robust to light conditions. S. Wang et al. * [16] proposed a new method of visible expression recognition by using the IR images data. It refined the SVM classifiers and deep models for both visible images and thermal images, by applying the restriction that the outputs of two views are similar for the SVM classifier. During the training the thermal were exploited for the construction of enhanced facial expression and representations classifiers from the visible images. They had extended the proposed method of thermal augmented expression recognition, because some time visible images and thermal images not recorded synchronously. MAHNOB laughter database was used for training and

testing. Proposed method's accuracy for posed laughter versus laughter discrimination was 1.57% and 17.50% higher than the SVM with only visible data. Proposed method had achieved up-to-the-mark expression recognition performance for both facial images unpaired and paired. The majority of the approaches used for object detection in thermal imaging assumed that object is considerably hotter than the background. Yan Zhang et al. * [22] presented a new algorithm for the detection of moving object due to which the main challenges of thermal image processing was overcome and got the precise silhouette shapes of objects and robust detection of foreground objects. Firstly background and foreground models were presented on the basis of thermal imagery properties. After that these models were used in unified MAP-MRF framework for the object detection. Two approaches were also compared quantitatively and qualitatively, two approaches were: GMM statistical background subtraction with the learning rate of 0.05 and the three Gaussian components and the Gaussian statistical background subtraction, which was the pre-step of the proposed method. Database used was sequence of thermal video that were recorded at the Ohio State University's campus at diverse environment. In term of precision the proposed method attained noticeable higher detection accuracy. Currently for fault diagnosis and reliable monitoring, cost effective and non-contact infrared thermo graphic system of inspection is used.

Study	**Year**	**Features**	**Technique**	**Database**	**Result**
E. Jackson et al. *[4]	2019	Otsu, k-means, adaptive thresholding and ATFD	(ATFD) Automated thermal footprint detector	Four environments of lab (sand, carpet, grass with foliage, and grass)	ATFD gave the highest score.
M. Haider et al. * [10]	2018	Skewness and Energy.	Threshold using edge detection. RGB image is converted into an HSV color model.	Total samples of the thermal images of the different electrical equipment were about 40.	Standard grey scale approach have more classification accuracy and have better feature extraction competence but proposed technique gave better results for

					all images as compared to it.
G. Lu et al. *[11]	2018	Edge orientation, Hu moment, Temperature feature, and Bag-of-words based on corner feature.	Feature Fusion	Xenics Gobi 640 GigE uncooled infrared camera was used for collecting the 3900 images.	90.1% accuracy was achieved.
C. K. Kyal et al. *[12]	2018	Accumulated Pixel Intensities and Vertical pixel location.	Preprocessing, Box Drawing and for Feature Extraction Coordinate Calculation	Dataset of more than 500 images were obtained in real time using FLIR 60101-0301 Model i7.	In a parallel environment the algorithm was much faster as compared to sequential environment processing time was 0.11.
S. Wang et al. * [16]	2018	*DNNs were used for feature extraction from both visible images and thermal images	Refined the SVM classifiers and deep models for both visible images and thermal images by applying the constraint that the output of two view are similar for the SVM classifier.	Database used was MAHNOB laughter.	Proposed method's accuracy for posed laughter versus laughter discrimination was 1.57% and 17.50% higher than the SVM with only visible data.
Jian-	2017	SIFT, BRIEF,	Receiver	Oxford image data	yBRIEF descriptor

Feng Shi et al. * [18]		ORB, sBRIEF AKAZE, yBRIEF and SURF	Operating Characteristics (ROC) metrics	set Graffiti and Boat	showed higher performance than the other BRIEF family descriptors
D. Zhou et al. *[19]	2011	N/A	HOG and SVM method was used as the pattern recognition method.	Images of the deer were captured in the Lake Superior Zoo in Duluth, MN USA.	The field and lab test results had showed 85% accuracy.
D. Kim et al. *[20]	2016	Histogram of Oriented Gradient (HOG) as a local shape feature	Temperature-based threshold method and shape features with a feature distance measure	Dataset OTCBVS was used.	Performed multi pedestrian tracking and detection effectively.
J. Hossen et al. * [21]	2015	Features are invariant under rotation, scale, translation, and reflection of images.	Universal segmentation and Feature Extraction technique	Data of real thermal image of repetitive human activity was recorded using FLIR A40 long-wave infrared.	76% Accuracy was achieved.
Yan Zhang et al. * [22]	2011	N/A	MAP-MRF decision framework	Sequences of thermal video that were recorded on the campus of Ohio State University at diverse environment.	In term of precision the proposed method attained noticeable higher detection accuracy.

Table 3.1: Literature Review Table for Conventional Methods

3.2 Convolutional and Deep Neural Networks:

Image classification has a lot of importance now days due to its large number of applications. Image classification and recognition is powered by the deep learning specifically CNN - Convolutional Neural Networks. Z. Jia et al. *[1] for the fault diagnosis of rotating machinery introduced a popular image feature extraction method that was based on IRT. Firstly had introduced the platform for experiment and then presented comparative experiments of two sets. For IRT images two most common feature extraction methods used were CNN and BoVW – bag of visual word and after feature extraction these are classified for implementation of automatic fault diagnosis. At the end classification of the extracted features was done to automate the fault diagnosis. Developed The experimental IRT images that was collected from the bearings was used to test the developed method and results demonstrated that developed method was more effective on the traditional method that was based on the signals of vibration Data was collected in two sets Group 1 and Group 2, each group contained 100 images. Results depicted that CNN have superiority in terms of the classification accuracy that was based on IRT images. G. Batchuluun et al. *[2] had proposed a human action recognition methods CycleGAN - cycle-consistent generative adversarial network by using image restoration, skeleton generation and halo effect removal , various ways was used to combine these approaches for generating different results. Techniques used for the implementation of proposed method were: CNN-LSTM, CNN and CycleGAN. (Dongguk activities & actions database (DA&A-DB2) an open database was used. Higher recognition rate was achieved ACC 99.6%, PPV 97.8% and TPR 98.3% than the existing methods. As a future work they will emphasis on the reduction or the removal of halo effects from the thermal images that was caused by the additional varied machines and objects in numerous environments. Furthermore, will work on using the lighter model with less parameters and with CNN-LSTM and CycleGAN for improving the processing time. S. Menon et al. *[7] conducted a research for sober or drunk classification and driver face recognition using Thermal Images. Face recognition of the driver was done by using a deep learning tool CNN and sober or drunk classification was done by using Fischer Linear Discriminant along with the Gaussian Mixture Model. The implemented process flow can be used on real time driver's photos and videos for their recognition as sober or drunk to control this misconduct at the earliest. Georgia Koukiou and Vassilis Anastassopoulos created the

database. 13% of the drunk drivers were classified successfully and 87% of the drivers were classified as sober. So the classification stage accuracy was almost 87%.Segmentation becomes very difficult for the human region if the human and background is at the same temperature. These difficulties affect the segmentation accuracy of the human region as well as reduce the human action recognition performance. E. Bartuzi et al. *[15] proposed a method of biometric recognition that was base on thermal images of the hand's inner part. For feature classification and selection two approaches were compared and proposed: Binarized Statistical Image Features (BSIF) was the feature engineering deploying texture descriptors, Gabor wavelets and feature learning used CNN that was trained in different environmental conditions. Two CNN architectures were used: SimpleNet and VGG-based CNN. The FLIR SC645 thermal sensor was used to acquire 21,000 thermal hand images for the study. 0.36% and 0.00% of EER - equal error rate was achieved in the first and the second approach. The study presented that hand temperature distributions are unique and recognition was close to perfect due to the proposed processing pipeline. Also presented that the hand thermal map's temporal stability is very limited. Ghenescu et al. *[17] presented an object detection method for thermal images of long range. This problem is extremely complicated because of the low resolution of long range thermal images, and also because of the fact that the objects need to detect had a very small area of 50 pixels. Presenting that the deep neural YOLO Darknet that is one of the fastest and powerful algorithm based on deep neural network can be used for the working on thermal images. Moreover, it was proved that objects that were hard to detect with the ordinary human eye was detected and classified by the network. Proprietary data set was used that include images one million captured by using thermal cameras of long range. Results that were produced by the best network had a score of 68.75%. Results that were obtained were very effective and will be able to add value to the existing surveillance systems of thermal cameras.

Study	**Year**	**Features**	**Technique**	**Database**	**Result**
Z. Jia et al. *[1]	2019	N/A	IRT image-based diagnosis scheme using BoVW and CNN.	Data was collected in two sets Group 1 and Group 2, each contained 100 images.	Results depicted that CNN have superiority in terms of the classification accuracy that was based on IRT images.
G.	2019	N/A	Techniques	(Dongguk	Higher recognition

Batch uluun et al. *[2]			used for the implementation of proposed method were: CNN-LSTM, CNN and CycleGAN	activities & actions database (DA&A-DB2) an open database	rate was achieved ACC 99.6%, PPV 97.8% and TPR 98.3% than the existing methods.
S. Menon et al. *[7]	2019	22 points on the face near the nose, chin area to emboss the junctures between capillaries and veins and near the eye brows.	CNN and Gaussian Mixture	Georgia Koukiou and Vassilis Anastassopoulos created the database.	Accuracy obtained was 87% approximately.
E. Bartuzi et al. *[15]	2018	Binarized Statistical Image Features	CNN Two CNN architectures were used: SimpleNet and VGG-based CNN.	The FLIR SC645 thermal sensor was used to acquire 21,000 thermal hand images for the study.	0.36% and 0.00% of EER - equal error rate was achieved.
V. Ghenescu et al. *[17]	2018	Boat, Tiny Boat, Human, Tiny Human, Animal, Tiny Animal, Vehicle, Tiny Vehicle,	DNN - YOLO Darknet	Proprietary data set was used that include images of one million captured by using thermal cameras of long range.	Results that were produced by the best network had a score of 68.75%.

Table 3.2: Literature Review Table for CNN

Some conventional methods did thermal image classification but very less work is done on the usage of deep learning for the thermal images classification. Models of the deep learning

achieve the highest accuracy; sometimes surpass the performance of humans. A deep learning algorithm convolutional neural network (CNN) is used for image classification. The requirement of pre-processing is much less for Convolutional Neural Network than the other classification algorithms. So, by keeping in view the accuracy of deep learning in the domain of image classification and benefits of thermal image classification, a research is conducted on developing a framework that is based on deep learning; work with thermal images that are captured by Seek Thermal and FLIR. Our own dataset is used no online dataset related to thermal objects human, car and cat is found. The proposed framework gives 93.31% accuracy for FLIR and 100% accuracy for Seek Thermal.

3.3 Deep learning:

Deep learning (DL) is the type of Artificial Intelligence and Machine learning mimic the way by which humans gain certain types of knowledge. In deep learning to perform the classification task, a model of the computer learns directly from the sound, images or text. Models of the deep learning achieve the highest accuracy; sometimes surpass the performance of humans [19]. Models of DL are trained on the large datasets and neural network architectures that include large number of layers. Deep learning plays an important role in data science, that contains predictive and statistics modeling.

Convolutional neural network is an algorithm of deep learning that is used for image classification, image recognition, object detection etc.; there are some other areas where CNNs are extensively used. It takes an input image, after that assign learnable weights to the numerous objects or aspects in the image (process it) and in end it classify object into certain categories – e.g. Cat, Human, Dog etc. In ConvNet the requirement of pre-processing is much less than the other classification algorithms. [20]
The main aim of ConvNet is the reduction of images in the format that make its processing easier, without any loss of critical features. For the designing of architecture that is good in learning features along with that scalable to huge datasets.

3.3.1 Convolutional Neural Network:

To form the CNN - Convolutional neural network architecture three layers are used: Pooling Layer, Convolutional Layer, and Fully-Connected Layer.

i. Convolutional Layer:

The first layer of CNN is Convolutional layer, it extract features from the input image. It learns the features of image in the form of small squares of the input data, in this way it preserves the relationship between pixels. It takes two inputs that are filter or kernel and image matrix after that generates the convolved output also known as “Feature Map”. In the figure 3 the 6 X 6 image matrix multiplies with the 3 X 3 kernel matrix and generates an output of 4 X 4 matrix. By convolutional of input image with different types of filters different operations are performed like blur, sharpen and edge detection.

Number of pixels that filter moves over the input matrix at a time is called stride. Filter moves to 1 pixel at a time if the stride is 1. The figure 3.1 shows the convolutional that work with a stride of 1.

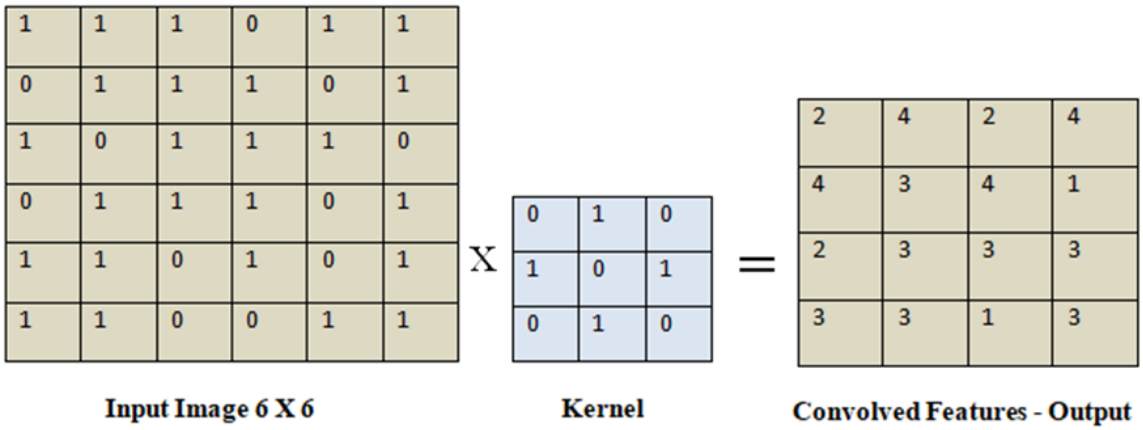


Figure 3.1: Convolutional of binary image

In the case of RGB images that have multiple channels. Matrix Multiplication is performed between In (RGB – I3) and Kn (RGB – K3) stack ([K1, I1]; [K2, I2]; [K3, I3]) and after it all the results are added with the bias to give a Convoluted Feature Output. In figure 3.2 an RGB is convolved by using three kernel channels and using bias = 1.

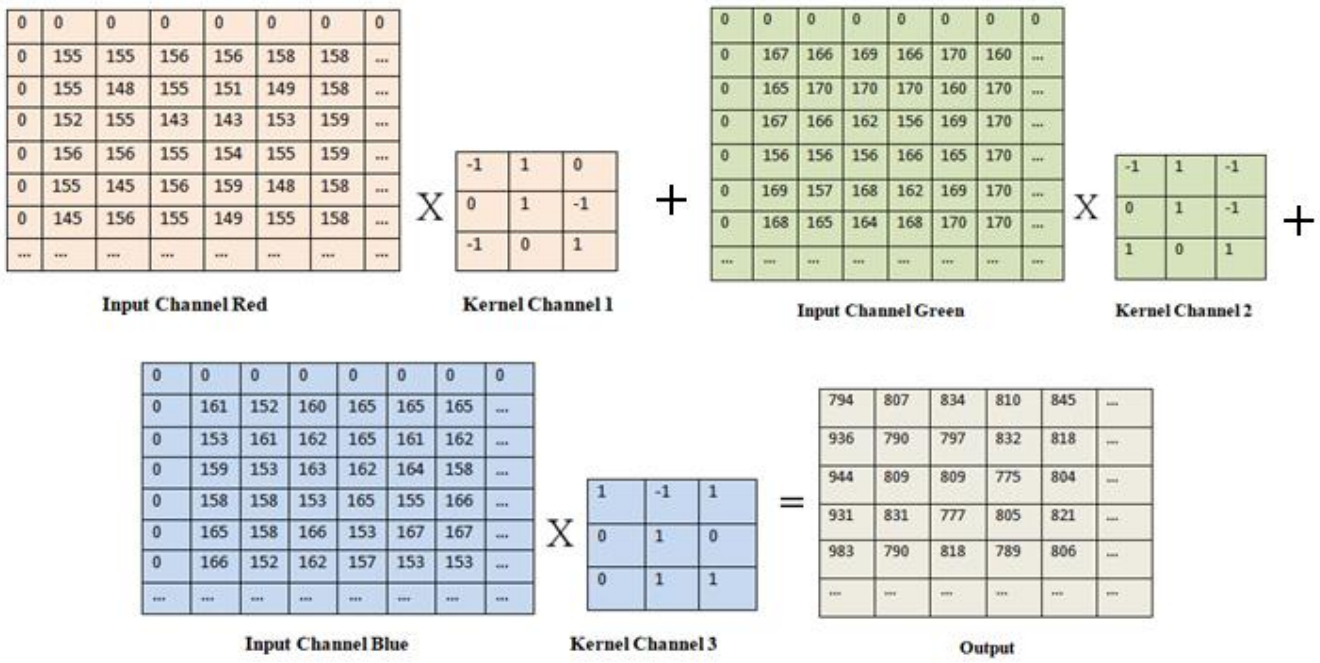


Figure 3.2: Convolutional of RGB image

ii. Pooling Layer:

To reduce the spatial size of the features the pooling layer is used that are convolved in the convolutional layer. By reducing spatial size computational power for data processing also get reduced. Moreover, it is also helpful in the extraction of dominant features, so that process of model training will get perform effectively.

Pooling is of different types like Average pooling, Global pooling and Max pooling, etc. The two most common pooling methods are the Average and the Max pooling. Average pooling returns the average of values that are covered by kernel in the image and Max pooling is the one that returns the maximum value from the values that are covered by kernel in the image. Max pooling remove the noise along with the reduction in the dimensions of the image. While average pooling performs the reduction in dimension and also adds smoothness in the image. Max pooling and the Average pooling is depicted in below figure 3.3.

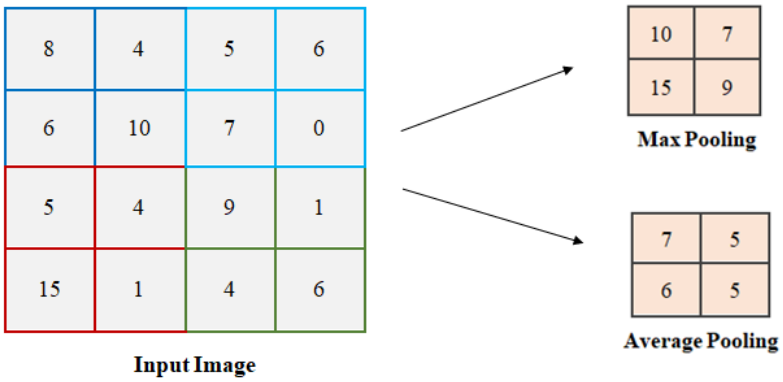


Figure 3.3: Max and Average Pooling

After completion of the above process the model will be able to understand the features. Afterward that needs to plane the final output and then feed that output to the commonly used Neural Network.

iii. FC Layer - Fully Connected Layer:

In this fully connected layer firstly there is need to flatten the matrix of the image into the vector and then feed that vector to the fully connected neural network.

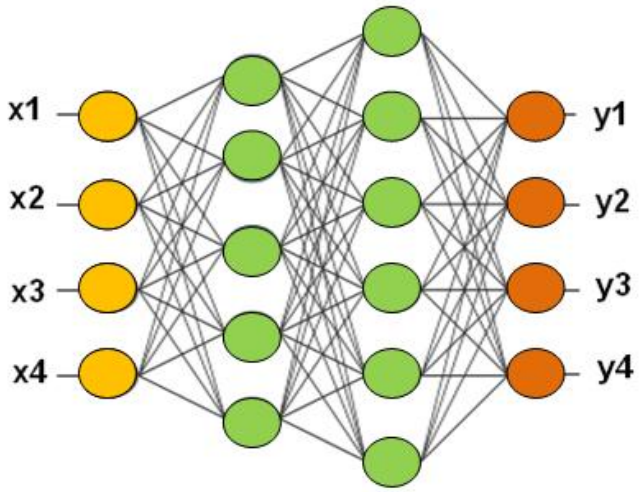


Figure 3.4: Fully Connected Layer

In the above figure 3.4 features [x1, x2, x3, and x4] are the inputs. Features are combined together for the creation of model. After that, over the number of epochs the model able to distinguish between high level and the low level features. Finally, the activation function like softmax is used for the classification of outputs like y1, y2, y3 and y4.

Figure 3.5 depicts the general architecture of the Convolutional Neural Network containing all layers.

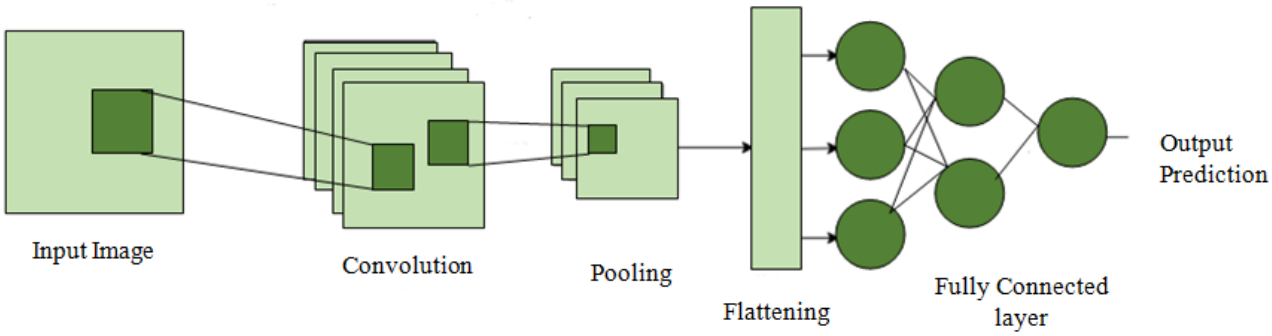


Figure 3.5: CNN Basic Architecture

There are different architectures of Convolutional Neural Networks are available which help in making the powerful AI algorithms. Some of the architectures are listed below:

- LeNet
- MobileNet
- GoogLeNet
- VGGNet
- ResNet
- DenseNet
- YOLOv4

CHAPTER 4: METHODOLOGY

This thesis presents method for the classification of thermal images. First we will explain the dataset, after that preprocessing and in end the classification methodology.

We have developed a framework based on deep learning; work with thermal images that are captured by using Seek Thermal and FLIR.

4.1 Database - Biomedical Image and Signal Analysis Lab, NUST CEME

Cameras we used to collect the training data are Seek Thermal and FLIR. Seek Thermal provides the thermal technology that was previously available only for the military and for some other professionals. It is a small camera that easily attaches to the Smartphone to get the thermal image of scenes around you; it shows a temperature shot of the environment. The Seek Thermal system contains Android or iOS device camera attachment and also provides an app that shows what the camera is capturing. Total 6414 were captured by using Seek Thermal. The image resolution is 300 x 400. The resolution of the images is higher in comparison with the handheld thermal cameras. The texture of the object is not clear but the temperature of the object is distinguishable, edges and contours are also clear.

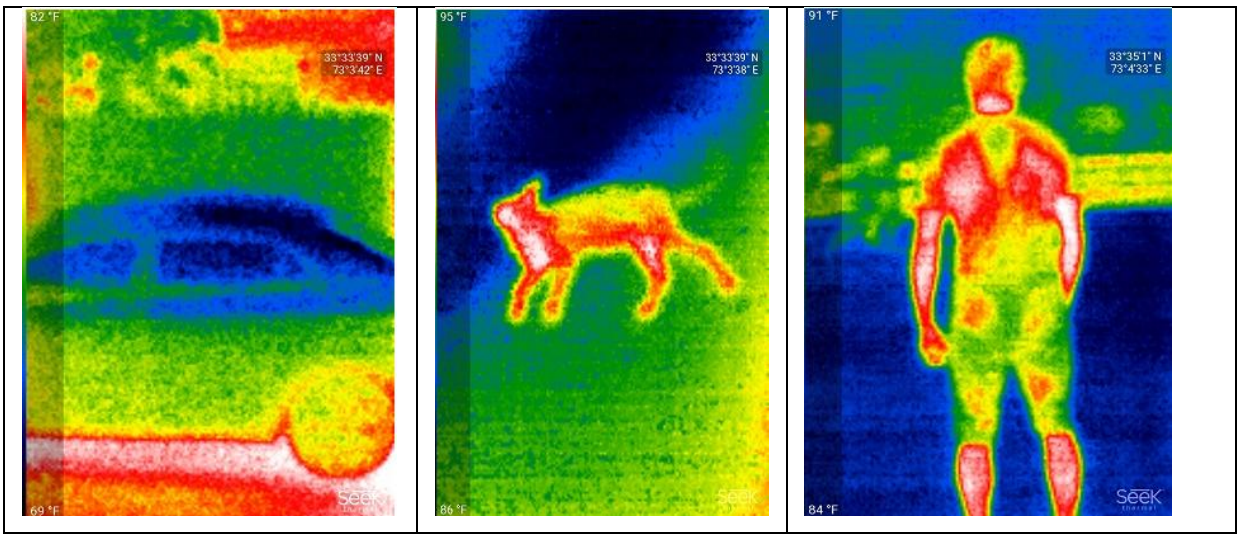


Figure 4.1: Sample images of car, cat and man captured by Seek Thermal

FLIR provides the faster identification of invisible problems whether the problem is related to the inspection of electric panel, looking into HVAC problems, object detection or troubleshooting mechanical systems. The camera FLIR ONE Pro offers the 4x resolution of

FLIR ONE Pro LT, for the image sharpening and making it clearer it is improved by the image processing of FLIR VividIR™ . Temperature measuring ability is also 3 x more than the other FLIR ONE model and have the higher sensitivity that allow it to measure temperature difference up to 70 mK. Total 1014 images are captured by using FLIR. The resolution of the images captured by FLIR camera is 1080 X 1440, which is higher as compared to image resolution of the images that are captured by SEEK thermal which we used for training the model.

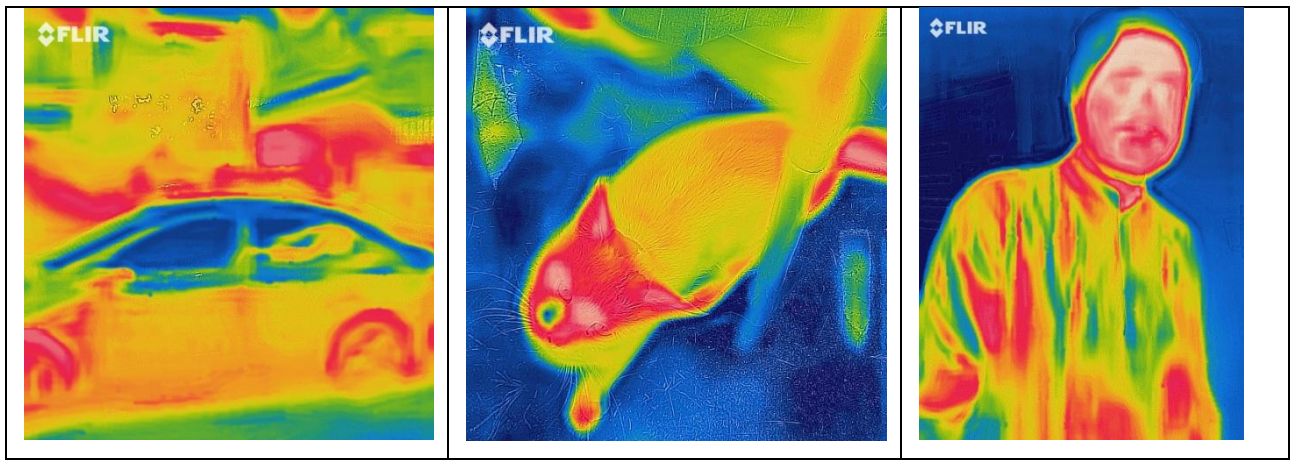


Figure 4.2: Sample images of car, cat and man captured by FLIR

Scanning	**Type**					
	FLIR			Seek Thermal		
Subject	Car	Cat	Man	Car	Cat	Man
Total Images	338	338	338	2138	2138	2138
Scan Resolution pixel x pixel	1080 x 1440	1080 x 1440	1080 x 1440	300 x 400	300 x 400	300 x 400

Table 4.1: BIOMISA Dataset

As shown in table 4.1 there are 3 subjects in both datasets FLIR and Seek Thermal respectively. In FLIR each subject contains 338 images with the resolution of 1080 X 1440 and in Seek Thermal each subject contains 2138 images with the resolution of 300 x 400.

4.2 Data Preprocessing

Before start tuning the model there is a need to prepare a data. Data contains images of cats, men and cars, which is further divided into training, and testing data. Data is stored on a disk in a particular directory structure or if using Colab then it is stored in the Google drive and need to mount the Google drive with the notebook. After this keras Image data generator is used to generate the data from the directory. Image Data Generator applies image transformation on the training images. Transformation includes zooming, rotation etc. By doing this new data get generated but the images generated by this are not completely different from original ones but provide variation of same data. This makes the model more robust.

For YOLOv4 model there was a need to label the images before training. Labels were assigned to all the images of Seek Thermal as well as FLIR. After assigning labels zip files were created and uploaded on the google drive for the training and testing of the model.

4.3 Classification

For thermal image classification CNN architecture was used. The DenseNet [30], MobileNet [27] and YOLOv4 [29] models are used for comparison and Inception v3 is used for classification after applying transfer learning. The last fully connected layer of the model was removed and then trained and tested on FLIR and Seek Thermal dataset. Figure 4.3 describes the methodology of classification. The dataset was randomly split into an 80:20 ratio as training and testing set respectively.

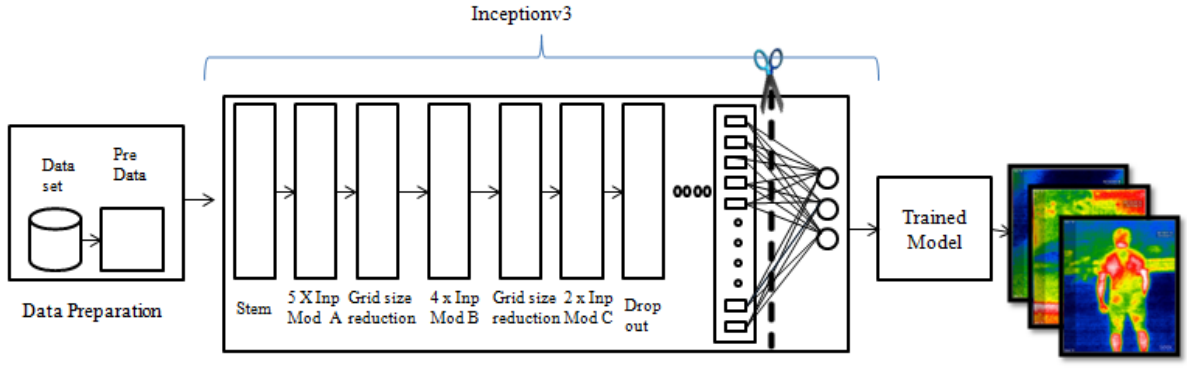


Figure 4.3: Classification methodology

The detailed structure of the model is discussed in the subsequent sections.

4.3.1 The Inception family

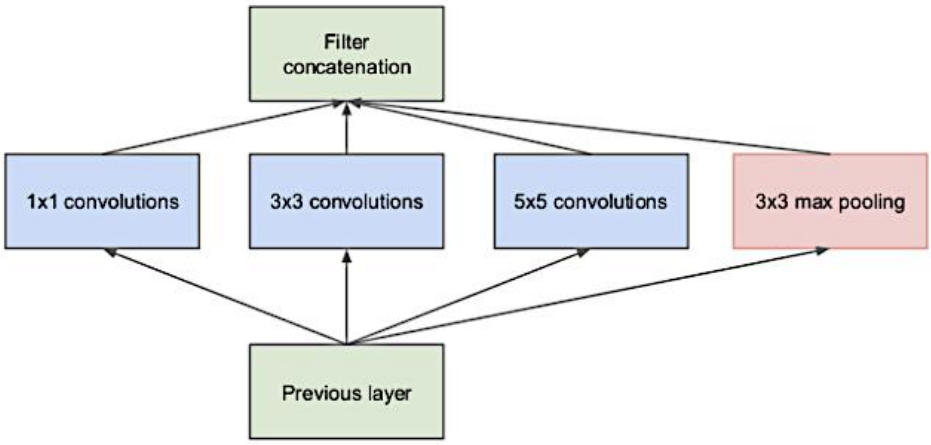
The inception family is going wider in contrast to the DenseNet models and ResNet model that was aimed at going deeper. From time to time for Inception model many versions were introduced, from them, we used Inception v3 for our framework. To understand the Inception v3 model there is a need to understand previous versions from where this model was evolved. Brief review of previous Inception model versions and Inception v3 is following:

Inception v1 (GoogleNet):

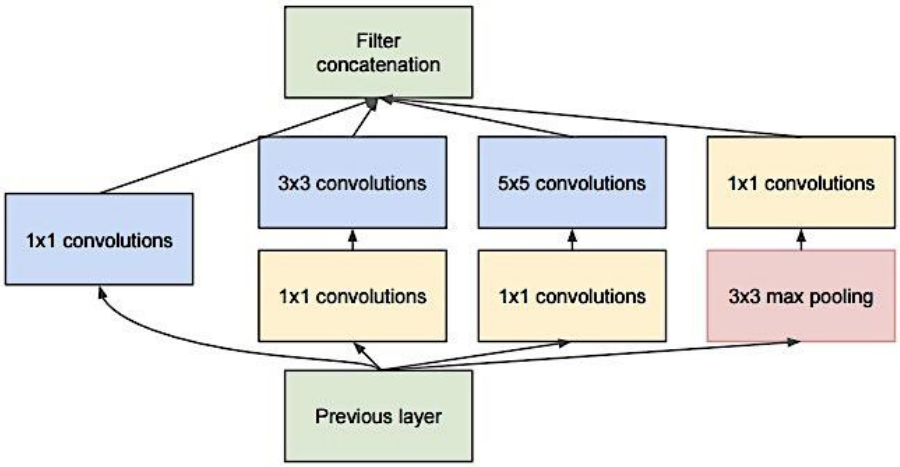
GoogleNet [*42*] is the first inception model, with its 6.7% top-5 error rate it is the winner of the ILSVRC 2014. As compared to previous network like VGG and AlexNet, with its 22 layers it is a much deeper network. However one of the main things about the GoogleNet is that it solve computational efficiency problem of training the larger nets. Without increasing the computational cost in GoogleNet they scaled up it by introducing the Inception module and then at top of each other a lot of these Inception modules were piled up. In this network there are no fully connected layers due to which total number of parameters gets reduced. In GoogleNet there are twelve times less parameters than AlexNet which had 60 million while GoogleNet has five million parameters.

In this network there is another network called Inception module shown in (Figure 4.16) that use varying filter size on the same input feature map that came from the previous layer to perform multiple convolutions in parallel and then results are concatenated. Convolutional transformations 5 x 5 and 3 x 3 are performed by the Inception module and in parallel a max-pool. For performing the varying size filter convolutions in parallel the main idea is that it permit the model to choose from among transformation that give the best information. The outputs of different filters give different information like output of 3x3 convolutional filter give different information from the convolutional 5 x 5 filter, and max pooling filter give different information, after applying all the operations and then concatenating them allow the model's subsequent layer to decide about the usage of piece of information. One of the major issues related to using this technique (stacking multiple different filters side by side) is that it increases the computational cost due to increase in the number of feature maps that develop into a lethal bottleneck in the model. This issue is catered by using the bottle neck layers, before the convolutional operations that is very expensive this bottle neck layer decrease the

input feature maps to lower dimension. To preserve the spatial dimension bottleneck layer performs 1x1 convolution that decrease the input depth by projecting the depth to a lower dimension like to compress the input of 64 x 64 x 100 to 64 x 64 x 20 twenty 1x1 filters are used. By reducing the input maps number , stacking different layer transformations enabled in parallel, that results in wide network (many operations in parallel) and deep (many layers).



(a)



(b)

Figure 4.4: (a) Inception module simple version
(b) Inception module with bottleneck layer[*42*]

Inception v2 and v3:

Following the Inception v1 model, few upgrades were proposed that reduce the computational cost and increase the accuracy, due to which different Inception versions came into being. Some of the major modifications in Inception v2 and v3 [47] are the following:

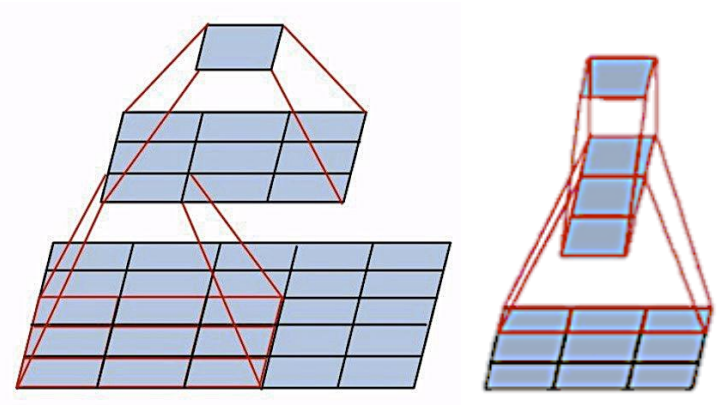
- Batch Normalization:

  All the feature maps at the output layer are normalized by using Batch-normalization in which output of layer is subtracted by its mean and then divide by its standard deviation that increase the network stability. This makes the same range for all the responses of neural maps and with mean zero, which corresponds to the data "whitening". Learning speed gets boosts up (by allowing higher learning rates) because it makes certain that no activation gone really low or high. Neurons that beforehand couldn't get to train will start to train by using this process. Slight regularization affect also introduces that helps in over-fitting reduction.

- Factorization:

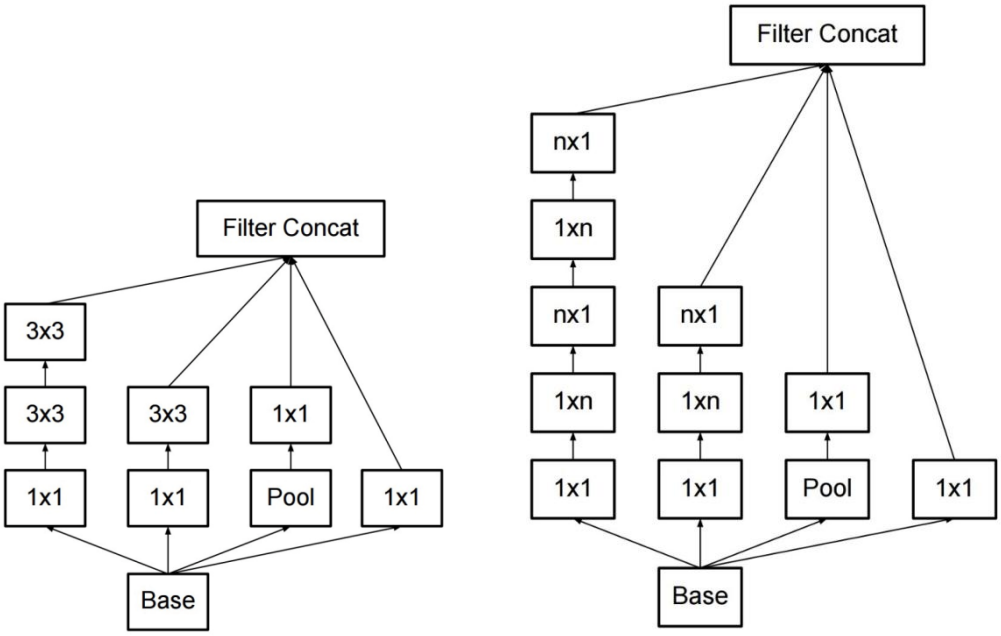
  Factorization is a method that is accelerating the computation, and reducing the number of parameters by splitting the larger convolutions into consecutive smaller ones without losing much expressiveness. Like, the 2 consecutive 3x3 convolutions replaced the 5x5 convolution (Figure 4.17(a)). By applying this technique, computations number was reduced to 18 (3*3+3*3) from 25 (5*5). After replacing the 5 x 5 convolutions the resultant Inception module is shown in Figure 4.18 (a).

  By using the asymmetric convolutions the filters can also be decomposed. It is depicted in the Figure 4.17 (b) that using 1 x 3 convolutions after 3 × 1 convolution is equivalent to the 3 × 3 convolution. So that in the theory, n x n convolution is replaced by using 1 x n convolutions after n x 1 convolution it reduce the cost of computation drastically as n grows Figure 4.18 (b). This factorization method performed well on medium sized grid (in range of 12 and 20) but practically it wasn't proved to be efficient on early layers.



(a) (b)

Figure 4.5: Bottleneck layer mini network replacing:
(a) 5x5 convolution (b) 3x3 convolution [47]



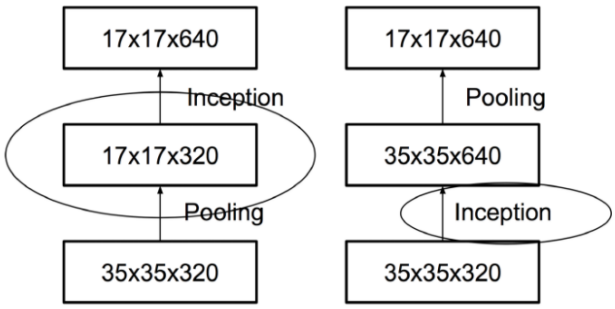
(a) (b)

Figure 4.6: Modified module of Inception by factorization:
(a) Symmetric convolutions where 5×5 convolution is replaced by two 3×3 convolutions (b) Asymmetric Convolutions [47]

- Grid Size Reduction

  Pooling operation was used by the CNN for the reduction of feature maps grid size that is generally followed by the operation of convolutional. Still on the larger grid the due to the expensive convolution the overall computational cost is dominated. One of the ways for minimizing the cost of computation is to switch the pooling layer with

the conv layer as depicted in Figure 4.19, but it make the information loss and make less significant network. So for inception a midway approach was used in which two blocks pooling and conv were used with 2 stride that reduced the grid-size efficiently while the filter bank's expanding as shown in Figure 4.20



(a) (b)

Figure 4.7: Two alternative ways of reducing the grid size

(a) Pooling followed by convolutions (representational bottleneck and loss of information) (b) Convolutions followed by pooling (expensive computations) [47]

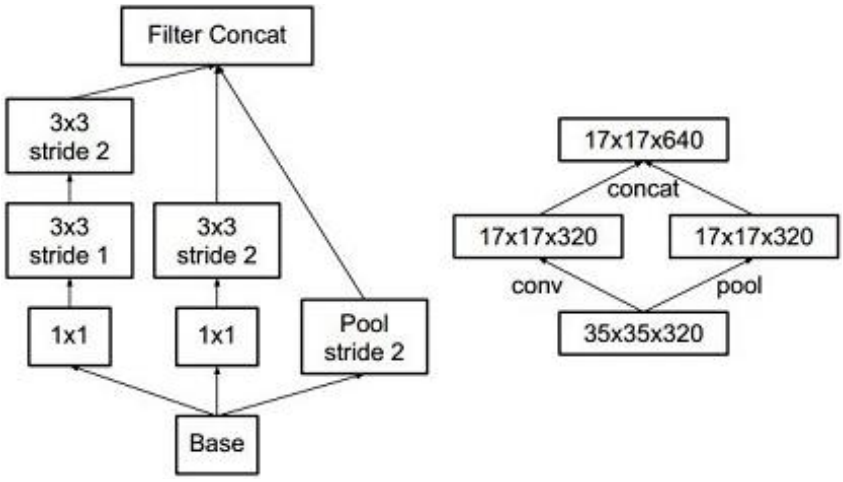


Figure 4.8: Modified Inception module with grid size reduction [47]

Inception v3 was used for the purpose of thermal image classification. In Figure 4.21 (a) entire configuration for ImageNet is outline. The last layer of the Inception v3 is modified according to number of classes desired.

Dataset	Epochs	Total Parameters	Trainable parameters	Non – Trainable parameters	Layers
Seek Thermal	12	23,851,784	23,817,352	34, 432	313
FLIR	30	23,851,784	23,817,352	34, 432	313

Table 4.2 Inception v3 parameters

4.3.2 Transfer Learning

In the transfer learning , firstly the base dataset is used to trained the network and task after that learned features are repurpose for the targeted network to be trained on a target task and dataset In practice, very less number of people train Convolutional neural Network from the scratch because it required a large dataset. It is a regular practice to train a CNN (Convolutional neural network) on a large dataset like ImageNet that contains million of images and then apply that model on some related different tasks after making a little modification in layers.

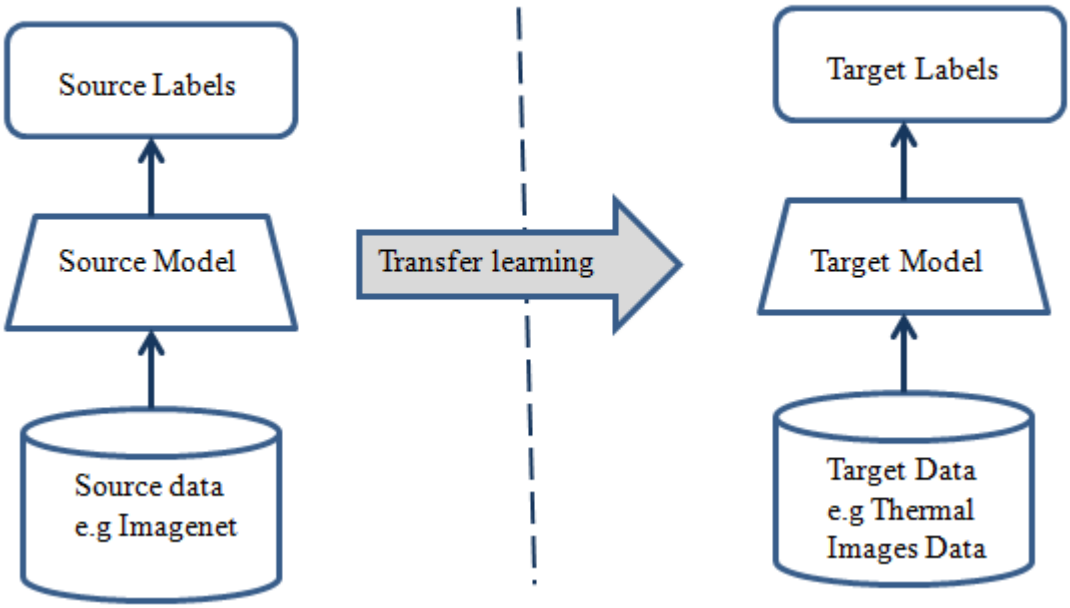


Figure 4.9: Transfer Learning

Transfer learning in computer vision is expressed as the utilization of pre trained models. Pre trained models are the models that were previously trained on large standard datasets (like the ImageNet) and solve the problem related to the problem that we want to solve. Training such models cost a lot due to which it is a very usual practice to import and use the model from the literature that is published (e.g. Mobile Net, DenseNet, Inception and YOLOv4).

i. Transfer learning process

The complete transfer learning process from practical point of view is summarized as follows:

1. First of all from a wide range of available models choose a pre trained model. Choose model that is most related to the problem that needs to be solved. Like if you are using Keras access InceptionV3 (Szegedy et al. 2015), VGG (Simonyan & Zisserman 2014).
2. After that remove the classifier that was originally used and adds classifier that fits your task.
3. In the end needs to fine tune the model, for that there are four different options:
    a. If the dataset set is different from the pre-trained model's dataset and are large in size. Train the model from the scratch because complete model training required large dataset.
    b. If the dataset is large and similar to the dataset on which model was trained previously. Then it is sufficient to train the top layers of convolutional base and the classifier, it will save time and huge effort of training.
    c. If the dataset is small and different from the dataset on which model was pre trained. Then there is need to find a balance between layers to freeze and to train. Model will get over fit if you will go deep and it will not learn anything if you stay shallow. In this case strategy mentioned in point 2 will help, that is train the top layers of convolutional base and the classifier.
    d. If the dataset is small and a problem that is similar to your problem was solved by the pre trained model. Then only last fully connected layer needs to be removed. As a fixed feature extractor, run the pre-trained model. After that, to train the new classifier use the resulting features.

Transfer learning is very beneficial because it speeds up the process of training model by reutilizing the modules or pieces of models that are already developed. It also accelerates the results.

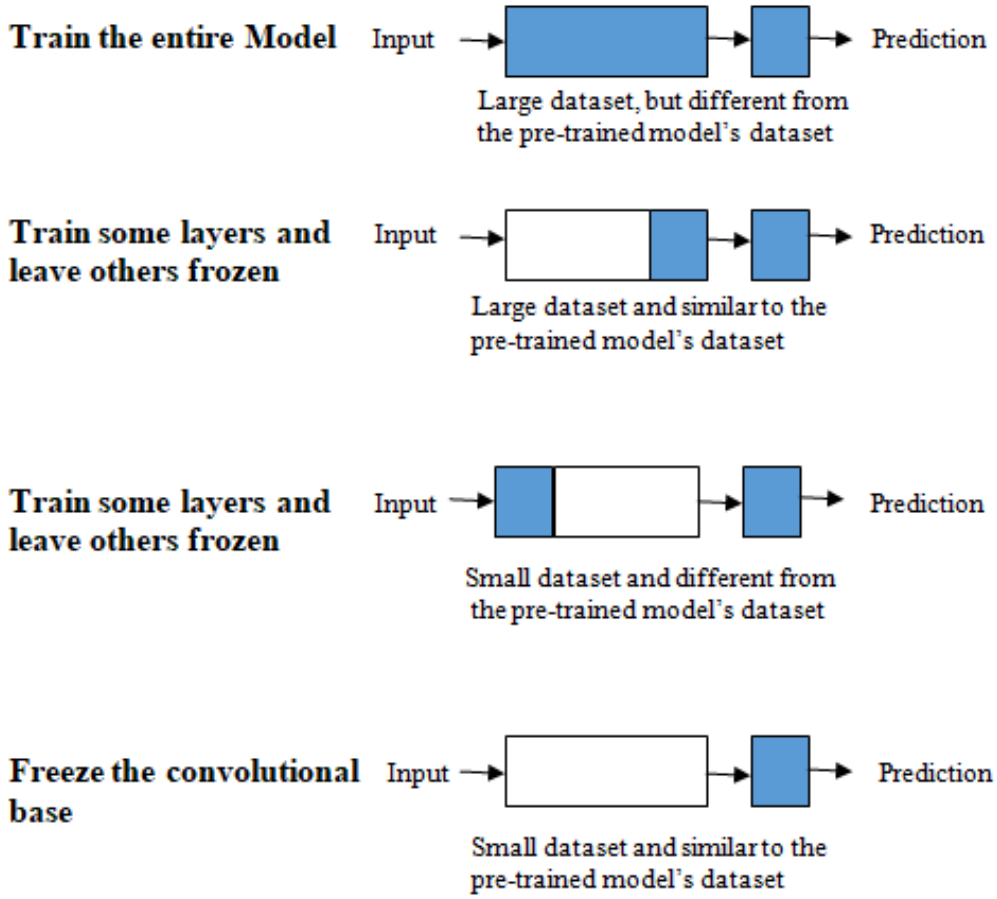


Figure 4.10: Fine tuning the model

4.4 Feature Maps and filters

Deep learning neural networks are mostly obscure, which means that they give very accurate predictions but why and how these predictions are made is not clear. CNN have structure that is designed to work with two dimensional image data. Model learns the two dimensional filters that can be visualized to find the type of features model is extracting.

Feature maps and filters for Inception v3 using Seek Thermal dataset and Inception v3 using FLIR dataset is presented in the figures below. Feature maps of colored image are compared with thermal images of cat, man and car.

Layer (Type)	Output Shape	Parameters	Connected To
conv2_block2_2_conv (Conv2D)	(None, 56, 56, 32)	36864	conv2_block2_1_relu[0][0]

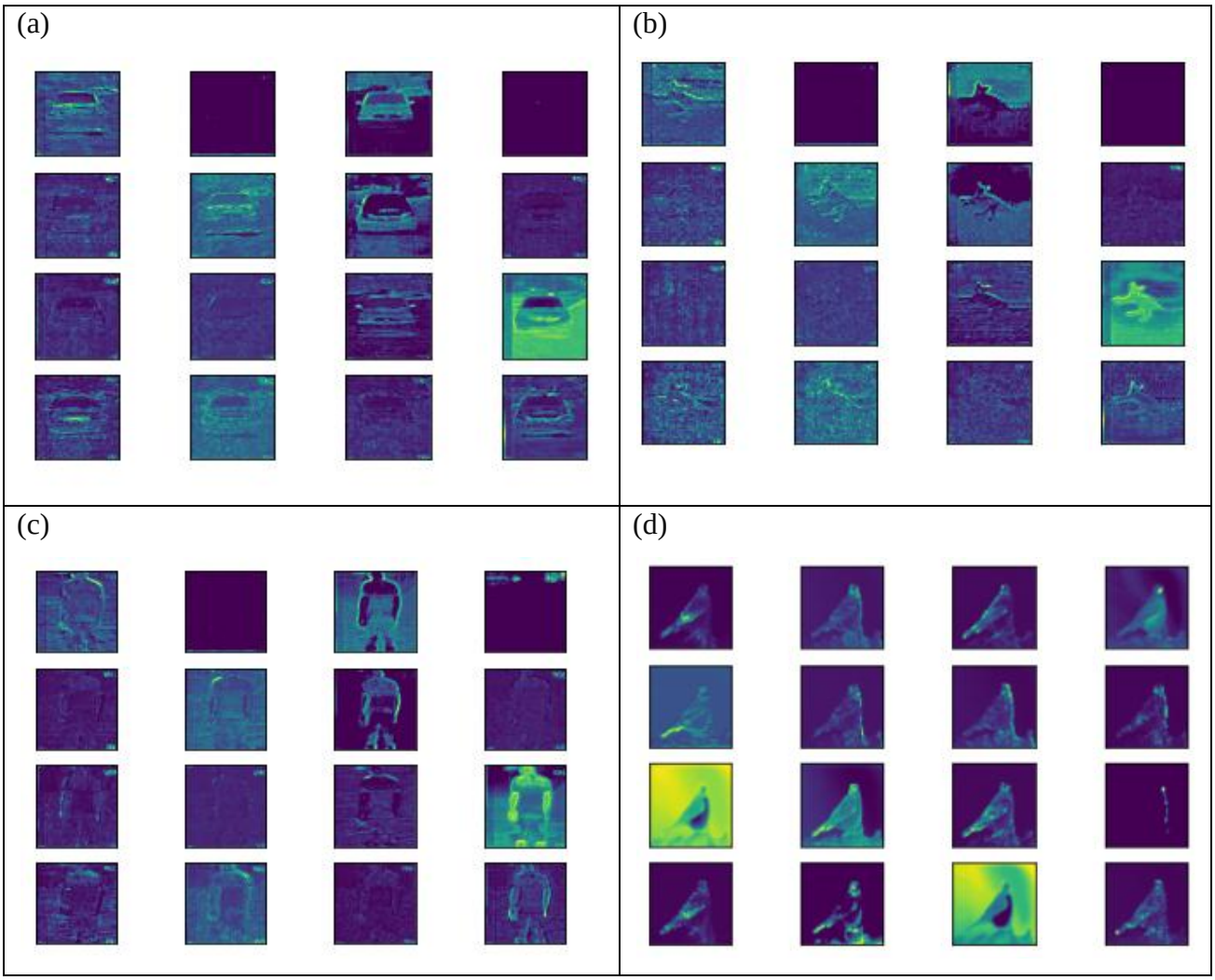


Figure 4.11: Feature map Inception v3 using Seek Thermal dataset – layer 10

(a) car (b) cat (c) man (d) colored image of bird

\

Layer (Type)	Output Shape	Parameters	Connected To
conv2_block2_2_conv (Conv2D)	(None, 56, 56, 32)	36864	conv2_block2_1_relu[0][0]

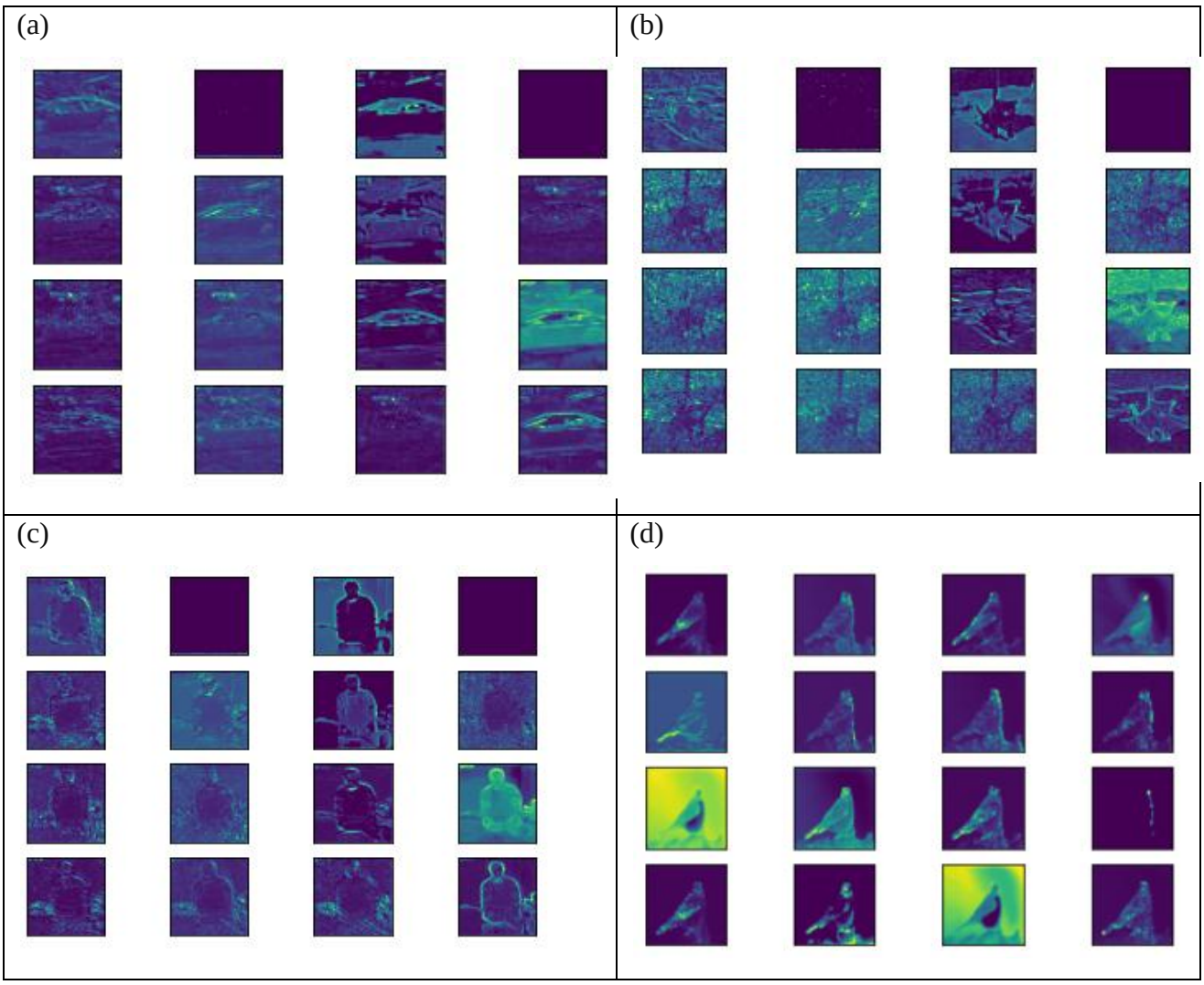


Figure 4.12: Feature map Inception v3 using FLIR dataset – layer 10

(a) car (b) cat (c) man (d) colored image of bird

Inception v3 model is used that consists of 313 layers and 23,851,784 parameters. Feature map for the layer 10 of Inception v3is presented in figure 4.21 for Seek Thermal and in figure 4.22 for FLIR. Feature maps of thermal images are compared with the feature maps extracted by Inception v3 for colored images. It is concluded that there is more sharpness in the images of feature map extracted by using colored images as compared to thermal images but still it give a very high accuracy.

CHAPTER 5: EXPERIMENTAL RESULTS

The proposed methodology's performance is evaluated on the local dataset provided by the Biomedical Image and Signal Analysis Lab, NUST CEME.

5.1 CNN Models used for Comparison

5.1.1 MobileNet

MobileNet has a higher accuracy with the fewer parameters. For further reduction in the parameters of network dense blocks were used that were proposed in DenseNet architecture [31]. Core layers of Mobilenet architecture are depicted in figure below.

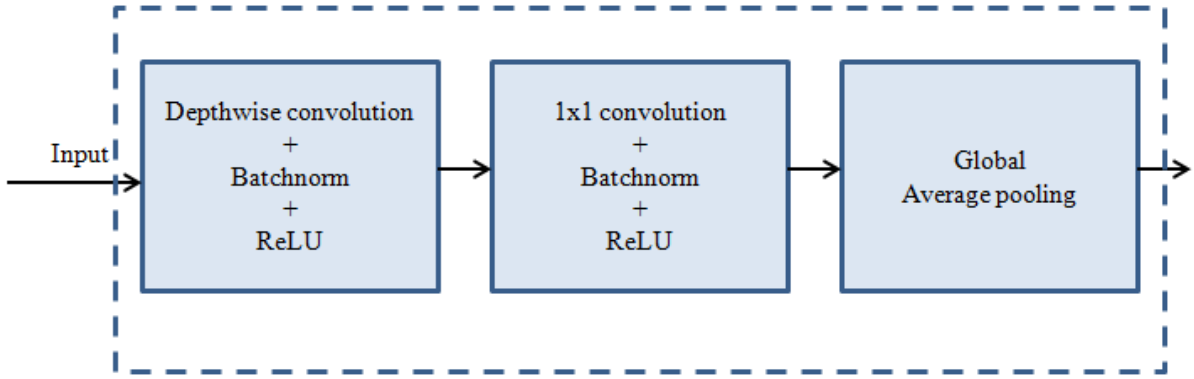


Figure 5.1: Mobile-net architecture [27]

Mobile Net architecture has 88 layers. In 88 layers of Mobilenet model there are 3×3 convolutional at the first layer following it there is 13 times the "depthwise separable convolutional" block. "Depthwise separable convolutional" block is divided into two layers: First is the depth wise convolutional layer that filter out the input, in which one input channel use one filter and then 1 X 1 point wise convolutional layer that merge the filtered values obtained to form the new features.

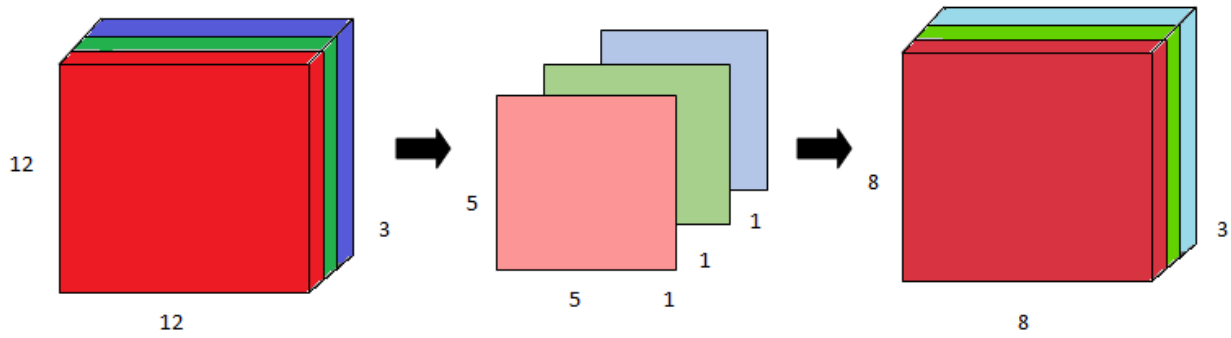


Figure 5.2: Depthwise Convolutional

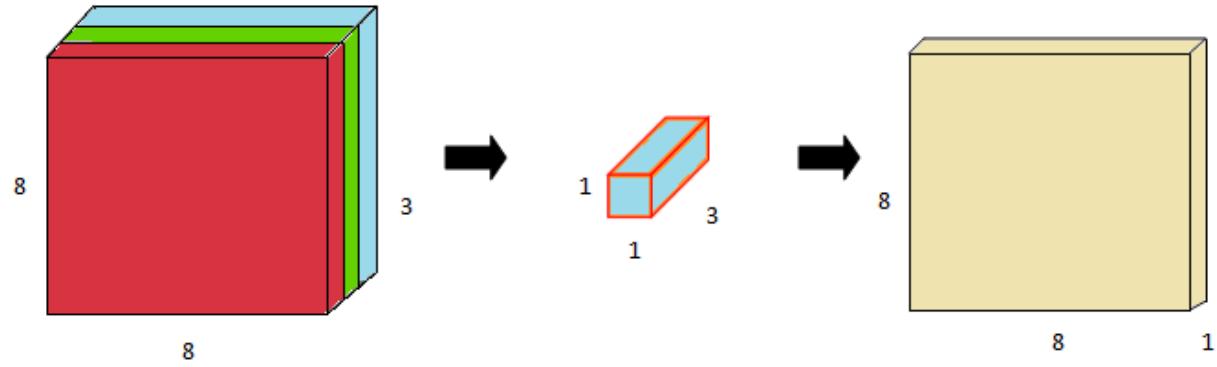


Figure 5.3: Pointwise Convolutional [24]

Depthwise convolutional in which one input channel used one filter could be written as:

$$O\hat{}_{k,l,m} = \sum_{i,j} F\hat{}_{i,j,m} \cdot I_{k+i-1,l+j-1,m} \qquad (5.1)$$

In it $F\hat{}$ is the kernel of Depthwise convolutional having the size of $S_k\, X\, S_k\, X\, P$ and the m_{th} filter is applied at the m_{th}channel in I, in the result of it for the filtered output feature map Oˆ the m_{th} channel is produced. The cost of the depthwise convolution is:

$$S_k . S_k . P . S_f . S_f \qquad (5.2)$$

In equation 2 the P represents the number of input channels, $S_k X S_k$ is the kernel size and $S_f X S_f$ the size of feature map. Relative to standard convolutional, depthwise convolutional is extremely efficient. It does not combine input channels to create new features, it filter the input channels only. The total cost of depthwise separable convolutional is:

$$S_k . S_k . P . S_f . S_f + P . Q . S_f . S_f \qquad (5.3)$$

The equation 3 is the addition of depthwise and 1 x 1 Pointwise convolutional. If represents the convolutional as the two step process then the computational cost will get reduce as:

$$\frac{S_k.S_k.P.S_f.S_f+P.Q.S_f.S_f}{S_k.S_k.P.Q.S_f.S_f} \quad (5.4)$$

$$= \frac{1}{Q} + \frac{1}{S_k^2} \quad (5.5)$$

In equation 4 the $S_k.S_k.P.Q.S_f.S_f$ represents the standard convolutional cost, in which Q represents the number of output channels.

Within the depthwise separable blocks there is no pooling layers. Instead of pooling layer some depthwise layers contain stride 2 that reduce the data's spatial dimensions, after that it's related Pointwise layer dual the output channels number e.g. If 224×224×3 is the input image then output will be of 7×7×1024 feature map. Batch normalization followed the Convolutional layers. In Batch normalization, normalization is the part of the architecture of model and for each training mini batch it provides normalization. It allows the usage of higher learning rates and initialization parameter also requires less care.

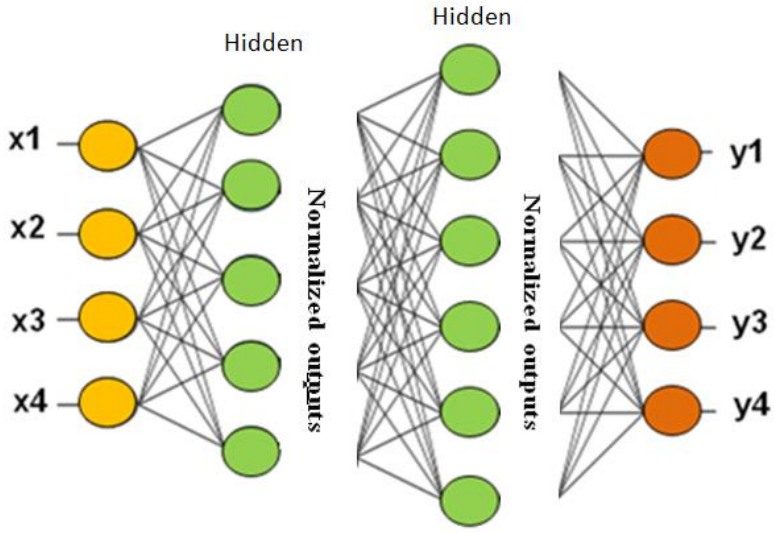


Figure 5.4: Batch Normalization [25]

To learn the complex patterns in the data, a function is added into the artificial neural network, that function is known as activation function. If compare the neuron based model with human brain then the activation function is function that decide what is essential to be

fired to next neuron. The same function it performs in the artificial neural network, it takes output from one layer and convert it into the form that can be the input of the next layer. The ReLU is simple calculation that directly gives the input as the value or if input is 0 or less directly give the value 0.

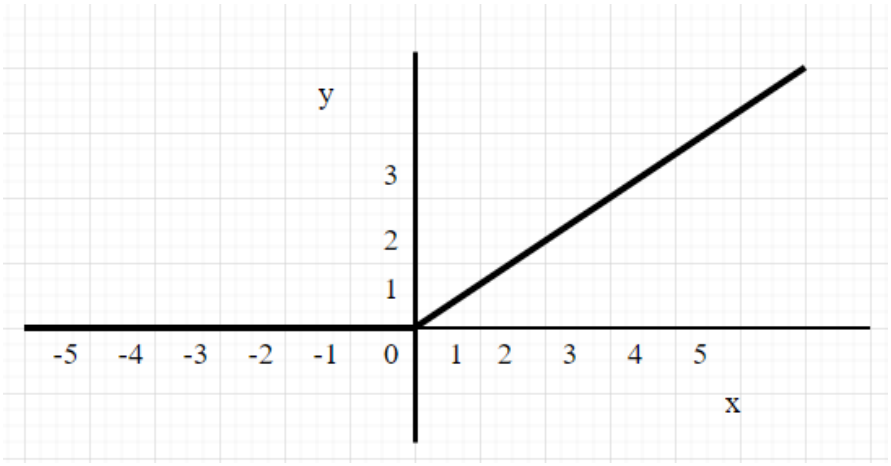


Figure 5.5: Activation function ReLU

MobileNet use the activation function ReLU6, it prevents the activation from getting too large. It restricts the ReLU at the positive side and it is represented as.

$$y = \min(\max(0, x), 6) \qquad (5.6)$$

It makes the function shape like a sigmoid function as sown in Figure 12. This restriction stops the activation from blowing by preventing the gradient to burst off -and also prevents from the other small issues that occur with ReLUs. It uses the low-precision computation due to which ReLU6 has been declared more robust as compare to ReLU. [21]

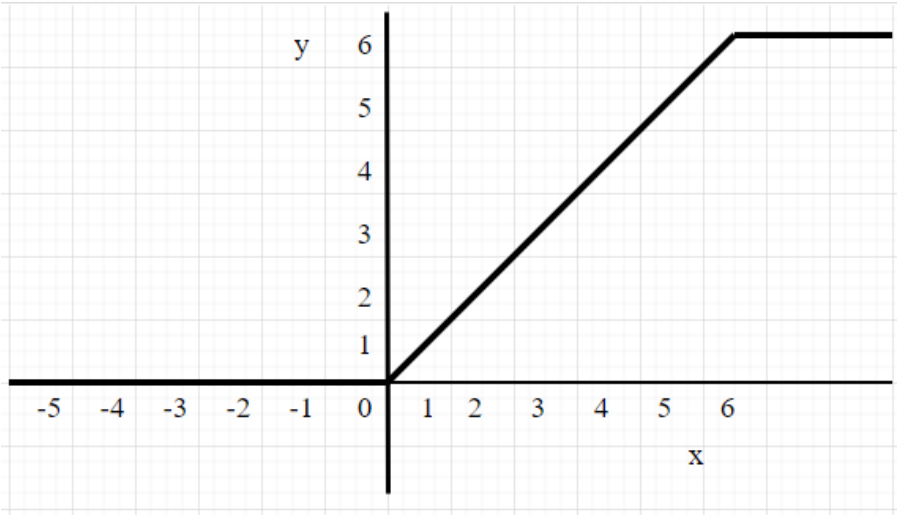


Figure 5.6: Activation function ReLU6

While using the low-precision computation ReLU6 is more robust as compare to ReLU. Based on the dataset and the scenario, the presented model uses a transfer learning process in which all layers of Mobilenet architecture are used excluding the last layer. After that an output layer is appended and then the activation function softmax is used.

5.1.2 YOLOv4

YOLO (You only look once) is accurate and fast and a family of detectors that detects in one stage. In a recent year, paper of YOLO v4 was released in comparison to the old object detectors it showed very good results.

In 2016, 2017 and 2018 the first three YOLO versions were released respectively. While in one year 2020, within few months YOLO's three versions were released that were named as YOLO v4, PP-YOLO and YOLO v5.

The latest Darknet based object detector version of YOLO is YOLOv4 for which a paper was also published with Alexey Bochkovskiy benchmarks[29]. However, for YOLOv5 no peer reviewed paper was published, till the time of writing this thesis.

Many GPUs are required for the training of modern accurate models with the large size mini-batch. It is impractical and slow to make the training by using one GPU. But YOLOv4 is an object detector that can be trained by using only one GPU, which resolve the issues faced by other modern object detectors. The state of the art results were achieved by using YOLOv4 at MS COCO dataset with AP 43.5% that is running at 65 FPS on V_{100} Tesla.

YOLO object detector has three different parts that are backbone, neck and head. Each part of new YOLOv4 is defined below.

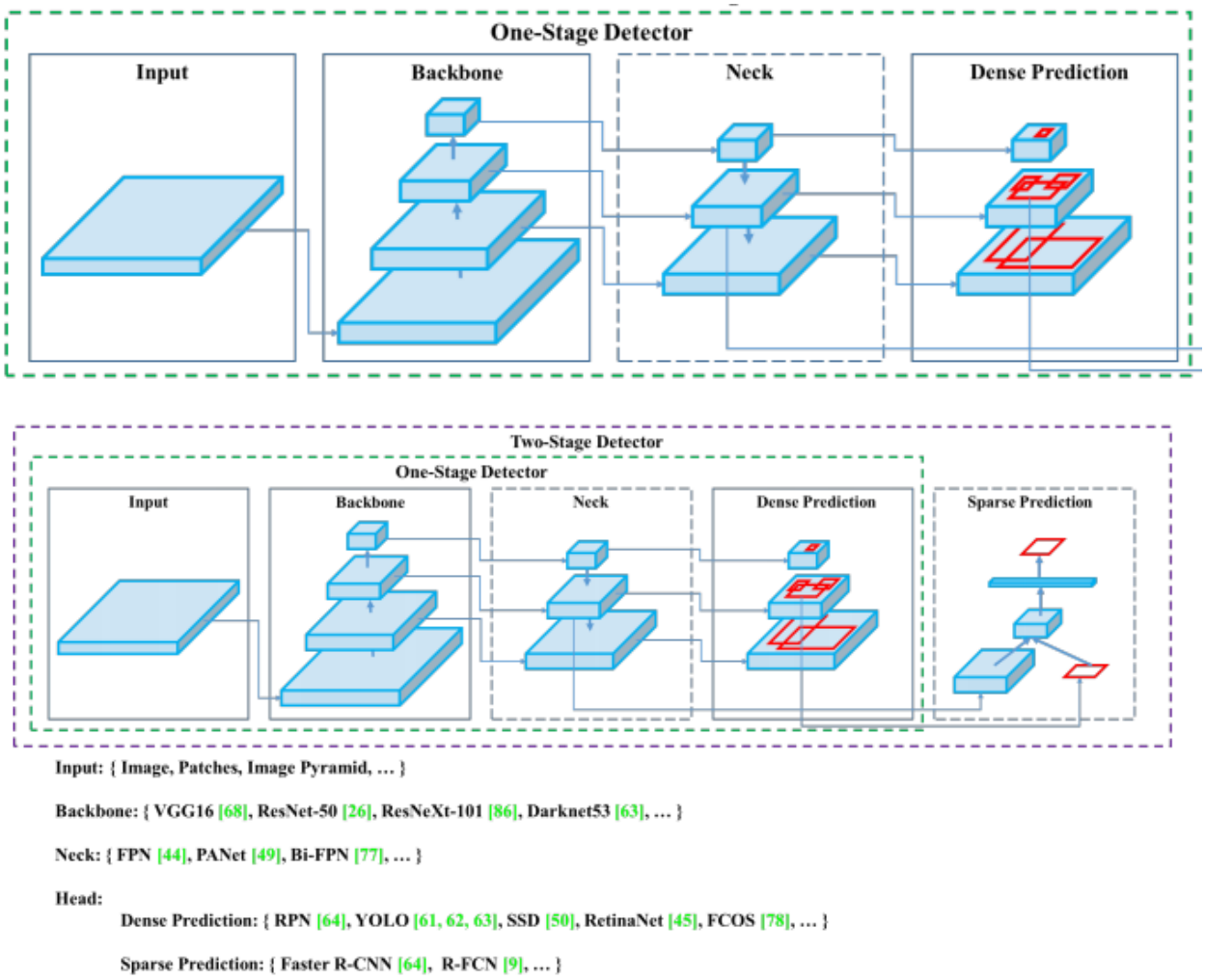


Figure 5.7: Object Detector [29]

i. Backbone:

For the GPU version YOLOv4 uses the feature extractor model CSPDarknet53. For Vision processing Unit (VPU) MobileNetV3, GhostNet, MixNet and EfficientNet-lite was considered using. Different backbones that were considered for GPU version are shown in following table.

Backbone model	Input network resolution	Receptive field size	Parameters	Average size of layer output (WxHxC)	BFLOPs (512x512 network resolution)	FPS (GPU RTX 2070)
CSPResNext50	512x512	425x425	20.6 M	**1058 K**	31 (15.5 FMA)	62
CSPDarknet53	512x512	725x725	**27.6 M**	950 K	52 (26.0 FMA)	**66**
EfficientNet-B3 (ours)	512x512	**1311x1311**	12.0 M	668 K	11 (5.5 FMA)	26

Figure 5.8: For the image classification parameters of neural networks [29]

Some of the backbones are more appropriate for detection and some are more suitable for classification. Like in terms of objects detection CSPDarknet53 is better than the CSPResNext50, while for the classification of image CSPResNext50 better than the CSPDarknet53. It is mentioned in the paper that, for best small object detection, higher input network size is required for a backbone model and for higher receptive field more layers are required.

ii. Neck:

Spatial pyramid pooling – SPP and PAN - Path Aggregation Network is used. PAN that is used in it is the tailored version not alike the original one. In it concat replaced the addition as shown in the figure 3.

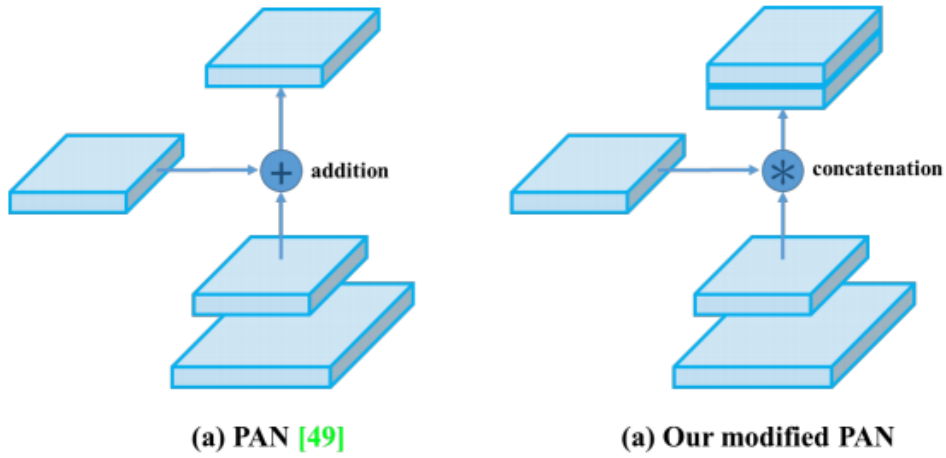


Figure 5.9: Modified PAN [29]

In original paper the PAN is formed by reducing the N4 size such that it have same spatial size like the P5 and then this new downsized N4 is added with P5 . At every level of Ni and Pi+1 this process is repeated for the production of N_i+1. In YOLO v4 concatenation is used for Ni and Pi+1 instead of addition. (Shown in above figure 4.12).

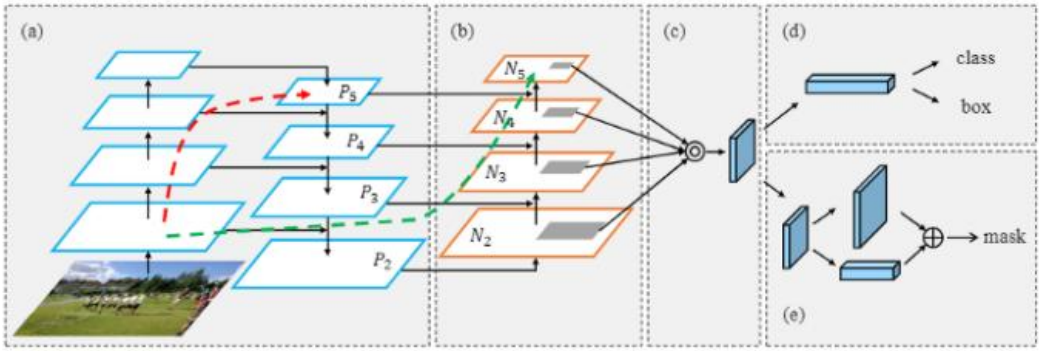


Figure 5.10: PAN (Path Aggregation Network)[29]

In **SPP** module using same padding (so that spatial size not get changed) and different kernel sizes k= {5, 9, 13} over 19*19*512 feature map max-pooling is performed. The volume of 19*19*2048 is formed by the concatenation of four feature maps. The neck receptive field gets increases with all this, which improve the accuracy of model with insignificant inference time increase.

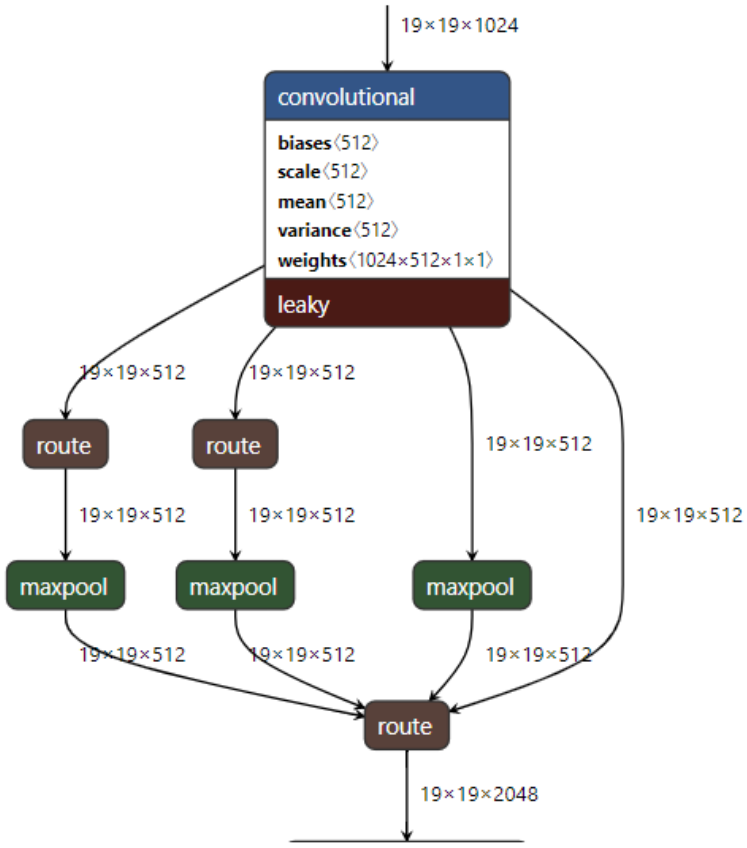


Figure 5.11: SPP observed in YOLOv4 [29]

iii. Head:

Same YOLOv3 head is used.

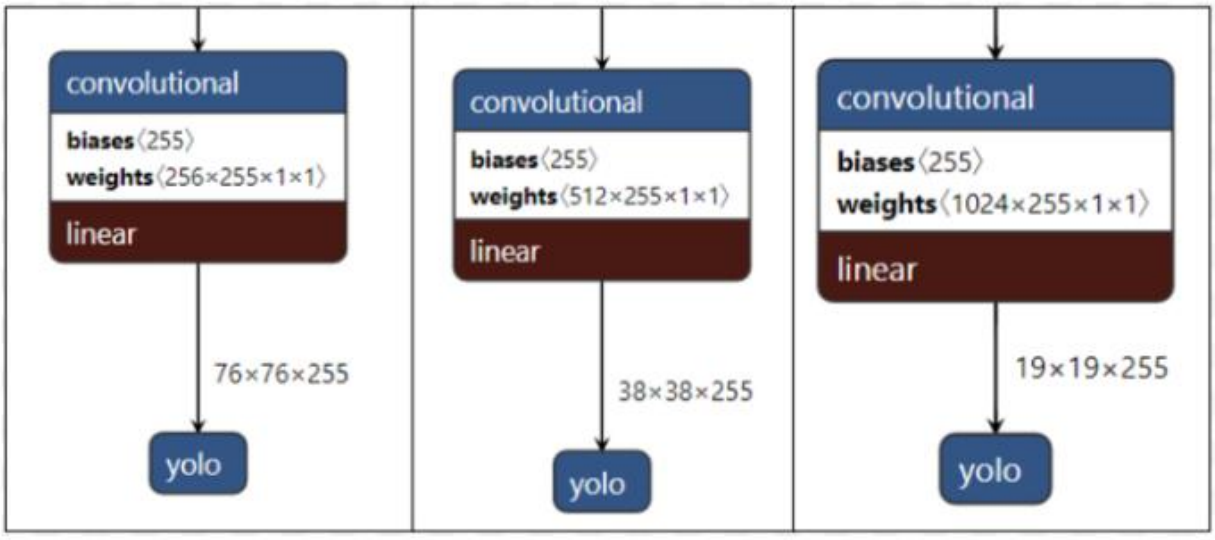


Figure 5.12: At different scales of the network the YOLO heads applied[29]

For the detection of different size objects the heads shown in the above figure are applied at different scales of the network. Total number of channels are 255 (80 *classes* + 4 *coordinates* + 1 *for objectness)* * 3 *anchors.*

The following table contains the summary of methods/modules of BoS and BoF that are used in the backbone as well as in the YOLOv4 detector.

	Backbone	Detector
Bag of Freebies (BoF)	• CutMix • Mosaic data augmentation • DropBlock • Class label smoothing	• CIoU-loss • Cross mini-Batch Normalization • DropBlock • Mosaic data augmentation • Self-Adversarial Training • Multiple anchors for a single ground truth • Cosine annealing scheduler • Optimal hyperparameters • Random training shapes
Bag of Specials (BoS)	• Mish activation • Cross-stage partial connections (CSP) • Multi-input weighted residual connections (MiWRC)	• Mish activation • SPP-block • SAM-block • PAN path-aggregation block • DIoU-NMS

Table 5.1: Summary of BoS and Bof[29]

iv. Additional Improvements:

For data augmentation ‘Mosaic’ a new method was introduced by the author of the paper. Idea that was used is to combine 4 training dataset images in 1. Using this new method on each layer, 4 different image’s activation statistics are calculated by Batch normalization due to which for training there is a requirement of selecting a large mini-batch size to reduce significantly. New method of augmentation is depicted in the images below..

Figure 5.13: Applying Mosaic

Self-Adversarial Training (SAT) was also used that works in 2 stages. In the stage 1 the original image was alter by the neural network instead of the network weights. By doing this adversarial attack is executes on the neural network itself, deception is created by altering the original image that in the image there is no desired object. In stage 2, for the detection of object in the modified image the neural network is trained.

5.1.3 Dense Network (DenseNet)

Recent work in the architecture of CNN has disclose that besides the direct connection of adjoining layers if add additional inter-layer connections then it helps in increasing the accuracy, depth and efficiency of training. By taking this observation in view Dense Convolutional Network was introduced that increased the information flow within the layers. In DenseNet each of the layer is linked to every other layer in the feed forward way, each layer get input from all preceding layers and this layer pass feature map of its own to the following layers. This layout is depicted in figure1. In comparison with the ResNet that

combine feature by using summation before passing it to the next layer, in DenseNet concatenation is used for combining features. In traditional neural network there are L connections between L layers, whereas in DenseNet the layer m^{th} has feature map inputs of all the preceding layers and feature map of its own are passed to all T- m. In this way $\frac{T(T+1)}{2}$ connections were introduced. Due to dense connectivity of this network it is known as DenseNet.

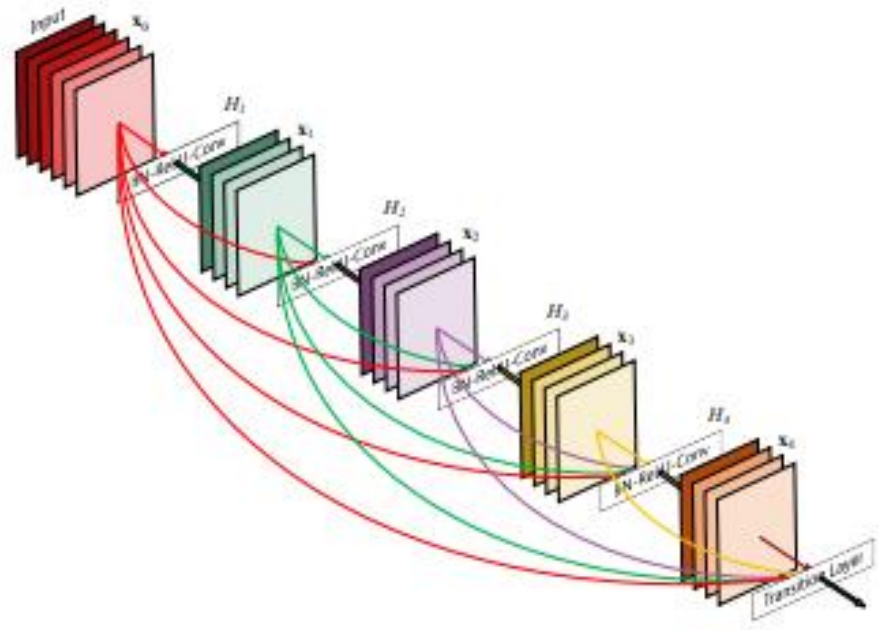


Figure 5.14: Dense block with 5 layers and the growth rate of k= 4[30]

DenseNet have many advantages.

1. There is no requirement of relearning the redundant feature map in DenseNet due to which DenseNet requires very less number of parameter as compare to other convolutional networks.
2. Layers of DenseNet are very narrow like per layer 12 filters that add a very little set of feature maps in network's complete knowledge while not making changes in the remaining feature maps.
3. One of the big advantages of DenseNet is their enhanced flow of gradients and information, due to which training of DenseNet is easy.

Consider an image x_0 that will pass in the convolution network. There are L layers in the network, each layer implements $H_m(.)$ which is non linear transformation , in it m is the layer index. For the operations like ReLU, Batch Normalization, Convolution or Pooling $H_m(.)$ can be a composite function.

i. ResNets:

Traditional convolutional networks join the m^{th} layer output as $(m + 1)^{th}$ layer input, that give the following transition:

$$x_m = H_m(x_m - 1). \qquad (5.7)$$

In ResNets, skip connection is added which bypass non linear transformations by an identity function, in it the output of H_l and identity function are combined by summation.

$$x_m = H_m(x_m - 1) + x_m - 1 \qquad (5.8)$$

ii. Dense connectivity:

For the improvement of information flow the direct connection from any of the layer to the entire subsequent layer were introduced in the DenseNet. The layout of DenseNet is illustrated in Figure1. Therefore, the layer m^{th} receives all preceding layers feature-maps, x0, . . . , x_{m-1}:

$$x_l = H_m([x_0, x_1, \ldots\ldots, x_{m-1}]) \qquad (5.9)$$

[x_0, x_1, . . . , x_{m-1}] is the feature maps concatenation of all the layers of network from 0, …m - 1.

iii. Composite function:

Three consecutive operations form a composite function that represented as $H_m(\cdot)$ in the DenseNet. Composite function contains: batch normalization (BN) [14], rectified linear unit (ReLU) [6] and a 3 × 3 convolution.

iv. Pooling layers:

The concatenation operation was used which is not useful if feature map size gets change. So, the down sampling layers are the most important part of Convolution networks because it changes the feature maps size. The architecture of DenseNet is divided into the multiple dense blocks as shown in figure 2. The layers between the dense block is known as transition layers that perform pooling and convolution.

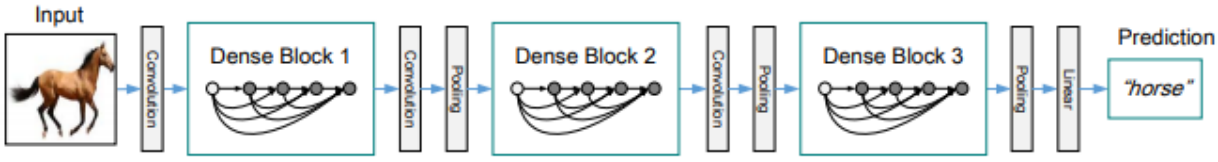


Figure 5.15: DenseNet containing three Dense Blocks.

v. Growth rate:

K features maps are produced by the each H_l function. Each of the dense layers receives input from all the previous layers, so the depth for m^{th} layer is:

$$D_m = k_0 + k * (m - 1) \qquad (5.9)$$

In it the k_0 represents the number of channels that are there in the input layer, DenseNet has very less number of layers as compare to the traditional architectures. The growth rate of the network represents by the K. View the global state of the network in terms of the feature maps. Every layer adds its own k feature maps to this state, here growth rate contributes and regulate that how much information every layer will add in the global state. In contrast with the traditional networks the global state that is written once can be accessed from all over the network.

vi. Bottleneck layers:

Each layer produce the output of k feature maps but still has many other inputs. In DenseNet a 1 X 1 convolution was introduced before 3 X 3 convolutions for the reduction of input feature maps this 1X 1 convolution is the bottleneck layer. By adding this layer computational efficiency got increase.

Table 1 contains the complete architecture of DenseNet 161, 121, 169 and 201 for ImageNet [30]. For thermal image classification we have used DenseNet 169 because of its better classification then others. The last softmax and fully connected layer are customized for our three classes. By using the Transfer learning the weights that were trained on ImageNet models were used.

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv, 3 × 3 conv] × 24	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 48	[1 × 1 conv, 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv, 3 × 3 conv] × 16	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Table 5.2: Complete DenseNet architectures for the ImageNet4.2.4

5.2 Performance Measures

For the evaluation of results common parameters are precision, recall, weighted average precision (WAP), weighted average recall (WAC), F1-score, misclassified and the accuracy (ACC). All the parameters are defined as:

$$Precision = \frac{Sum\ of\ TP}{Sum\ of\ TP + Sum\ of\ FP} \tag{5.10}$$

$$Recall = \frac{Sum\ of\ TP}{Sum\ of\ TP + Sum\ of\ FN} \tag{5.11}$$

$$\begin{aligned} WAP = {} & Actual\ class\ car\ instances\ x\ precision\ of\ class\ car \\ & + Actual\ class\ cat\ instances\ x\ precision\ of\ class\ cat \\ & + Actual\ class\ man\ instances\ X\ precision\ of\ class\ man \end{aligned} \tag{5.12}$$

$$\begin{aligned} WAR = {} & Actual\ class\ car\ instances\ x\ recall\ of\ class\ car \\ & + Actual\ class\ cat\ instances\ x\ recall\ of\ class\ cat \\ & + Actual\ class\ man\ instances\ x\ recall\ of\ class\ man \end{aligned} \tag{5.13}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.14}$$

$$F1 - score\ = \frac{2\ x\ precision\ x\ recall}{precision\ x\ recall} \tag{5.15}$$

For classification, TP (True Positive) for cat, man and car class is the no. of cat, man or car correctly classified as cat, man or car. FP (False Positive) for cat, man and car class is the no. of cat, man or car wrongly classified as cat, man or car. TN (True Negative) for cat is the no. of man correctly classified as man or no. of car correctly classified as car. TN (True Negative) for car is the no. of man correctly classified as man or no. of cat correctly classified as cat. TN (True Negative) for man is the no. of car correctly classified as car or no. of cat correctly classified as cat. FN (False Negative) for car the no. of man wrongly classified as man or no. of cat wrongly classified as cat. FN (False Negative) for car is the no. of man wrongly classified as man or no. of cat wrongly classified as cat. FN (False Negative) for man is the no. of car wrongly classified as car or no. of cat wrongly classified as cat. Following matrix shows the TP, TN, FP and FN for the cat class.

	Cat	Car	Man
Cat	TP	FN	FN
Car	FP	TN	TN
Man	FP	TN	TN

Table 5.3: Matrix showing TP, TN, FP and FN for the cat class

5.3 Results

5.3.1 Results of Classification

Classification methodology was evaluated on Seek Thermal dataset. The dataset was split into 80:20 (≈5346, 1068) randomly as training and testing data respectively. There are three classes in the input data, cat class, car class and man class. Table 5.2 demonstrates the results of proposed framework.

Database	**Model**	**No. of Images**	**Training (80%)**	**Testing (20%)**	**Prec %**	**Recall %**	**WAP %**	**WAR %**	**ACC %**	**F1-Score**	**Miss Classified %**
FLIR	Inceptionv3	1014	807	207	98.90	98.90	99	97	98.91	0.98	1.09
Seek Thermal	Inceptionv3	6414	5346	1068	100	100	100	100	100	0.100	0

Table 5.4: Results of proposed framework

Table 5.3 demonstrates the comparison of proposed framework results in term of accuracy with other models. Column 2 of the table 5.3 shows accuracy of models for the FLIR dataset and column 3 of the table 5.3 shows the accuracy of models for the Seek Thermal dataset.

Model	**ACC (%) for FLIR**	**ACC (%) for Seek Thermal**
Inception v3	98.91	100
DenseNet (D)	93.31	89.61
MobileNet (M)	86.19	100
YOLOv4	84.63	85.53

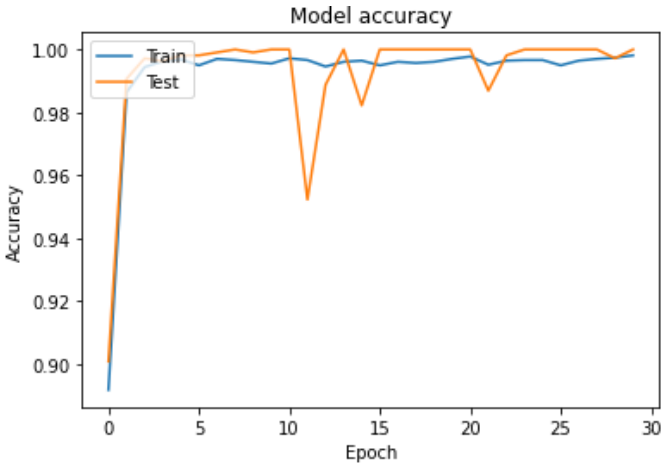
Table 5.5: Comparison of results in terms of accuracy

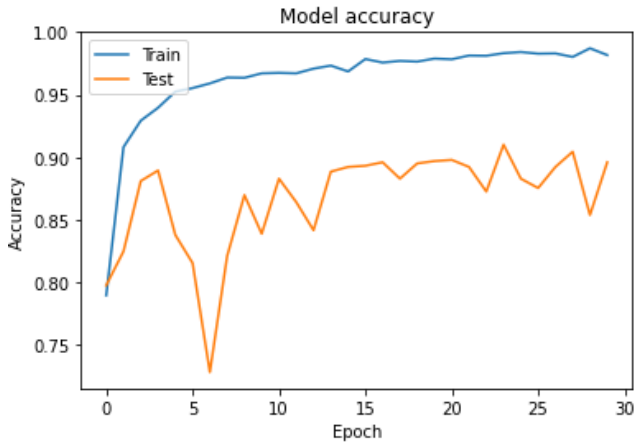
Table 5.4 shows the comparison of different models DenseNet, MobileNet and YOLOv4 with proposed framework. These models are also trained and tested using the same dataset on which proposed framework is trained and tested. Comparison is performed on basis of precision, recall, weighted average precision (WAP), weighted average recall (WAC), F1-score, misclassified and the accuracy (ACC) of individual models for FLIR and Seek Thermal dataset.

Database	**Model**	**No. of Images**	**Training (80%)**	**Testing (20%)**	**Prec %**	**Recall %**	**WAP %**	**WAR %**	**ACC %**	**F1-Score**	**Miss Classified %**
FLIR	Inceptionv3	1014	807	207	98.90	98.90	99	97	98.91	0.98	1.09
	YOLOv4				74	83	74	83	84.63	0.78	15.37
	DenseNet				93.30	93.39	93.97	93.17	93.31	0.93	6.69
	MobileNet				86	86	87.29	85.86	86.19	0.86	13.81
Seek Thermal	Inceptionv3	6414	5346	1068	100	100	100	100	100	0.100	0
	YOLOv4				87	84	87	84	85.53	0.85	14.47
	DenseNet				89.60	89.60	90	89	89.61	0.89	10.39
	MobileNet				100	100	100	100	100	0.100	0

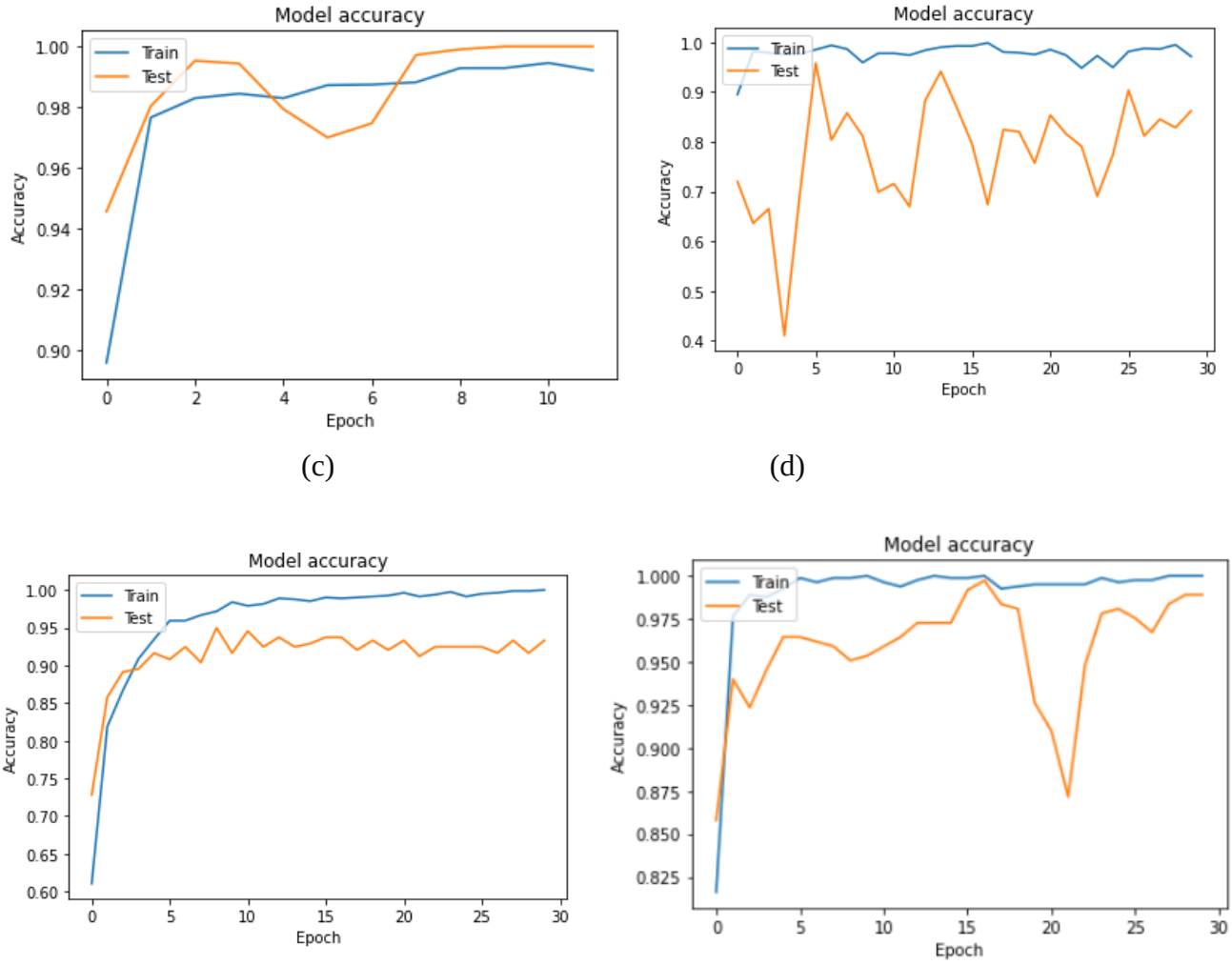
Table 5.6: CNN Models Comparison with proposed framework

The performance of the models DenseNet, MobileNet and Inception v3 for FLIR and Seek Thermal are depicted in the graphs below; In the graphs the blue lines are depicting the training accuracy and the orange lines are depicting the testing accuracy, the lines of the graph clearly shows that models are not under fit as well as not over fit. Testing accuracy is increasing with the number of Epoch. But as shown in graph with further increase in Epoch the models will get over fit.





(a) (b)



(c) (d)

(e) (f)

Figure 5.16: Training and testing accuracy trend with respect to no. of epochs of (a) MobileNet using Seek Thermal dataset (b) DenseNet using Seek Thermal dataset (c) Inceptionv3 using Seek Thermal (d) MobileNet using FLIR dataset (e) DenseNet using FLIR dataset (f) Inceptionv3 using FLIR dataset

Following figure contains the confusion matrix for Inceptionv3 using Seek Thermal, mobile Net using Seek Thermal dataset and DenseNet using Seek Thermal dataset.

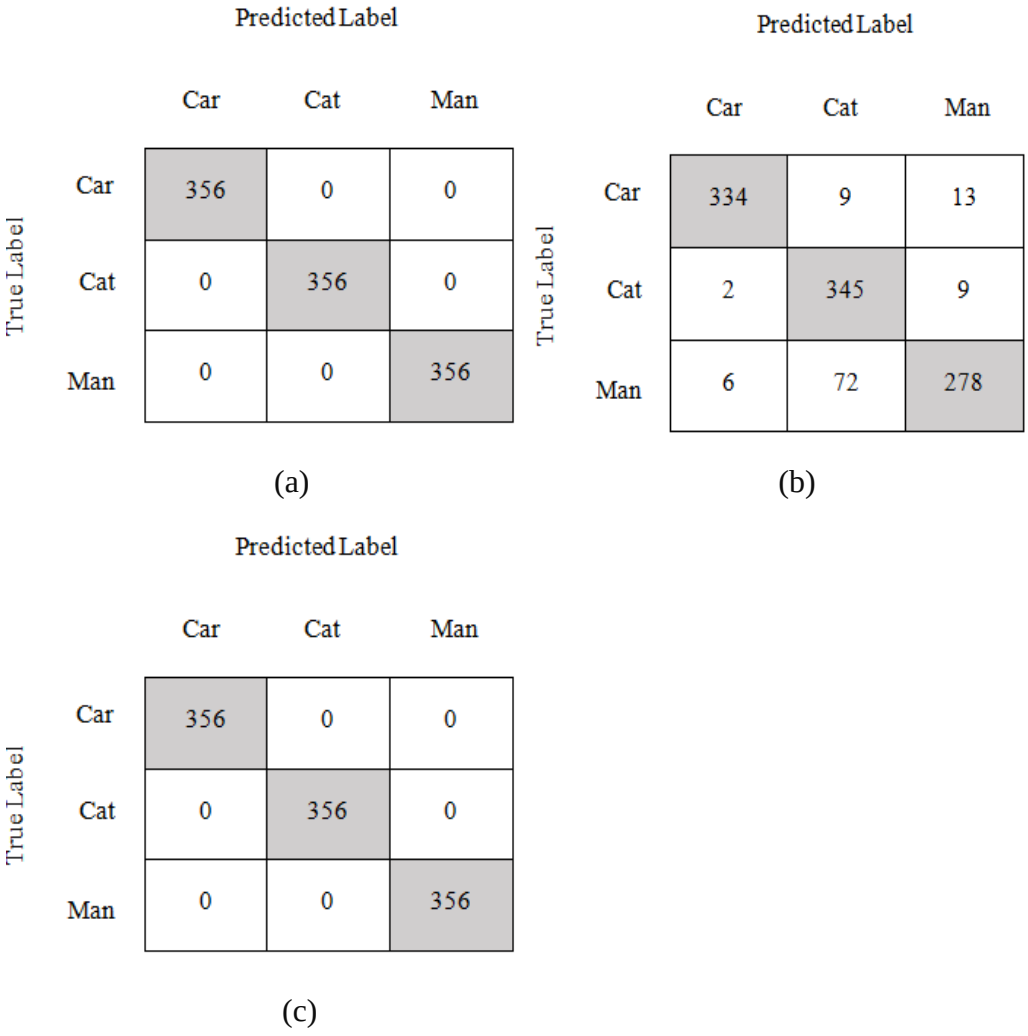


Figure 5.17: Confusion Matrix for Seek Thermal
(a) Inceptionv3 (b) MobileNet (c) DenseNet

Confusion matrix given in Figure 5.2 (a) depicts that for car class, 356 cars are predicted correctly while 0 cars are predicted as cat and 0 cars are predicted as man, for cat class 356 cats are predicted as cats, 0 cat is predicted as car and 0 cats are predicted as man and for man class 356 men are predicted as man, 0 men are predicted as cat and 0 men are predicted as car. Confusion matrix given in Figure 5.2 (b) depicts that for car class, 334 cars are predicted correctly while 9 cars are predicted as cat and 13 cars are predicted as man, for cat class 345 cats are predicted as cats, 2 cats are predicted as car and 9 cats are predicted as man and For man class, 278 men are predicted as man, 72 men are predicted as cat and 6 men are predicted as car. Confusion matrix given in Figure 5.2 (c) depicts that for car class, 356 cars are predicted correctly while 0 cars are predicted as cat and 0 cars are predicted as man, for cat class 356 cats are predicted as cats, 0 cat is predicted as car and 0 cats are predicted as

man and for man class 356 men are predicted as man, 0 men are predicted as cat and 0 men are predicted as car.

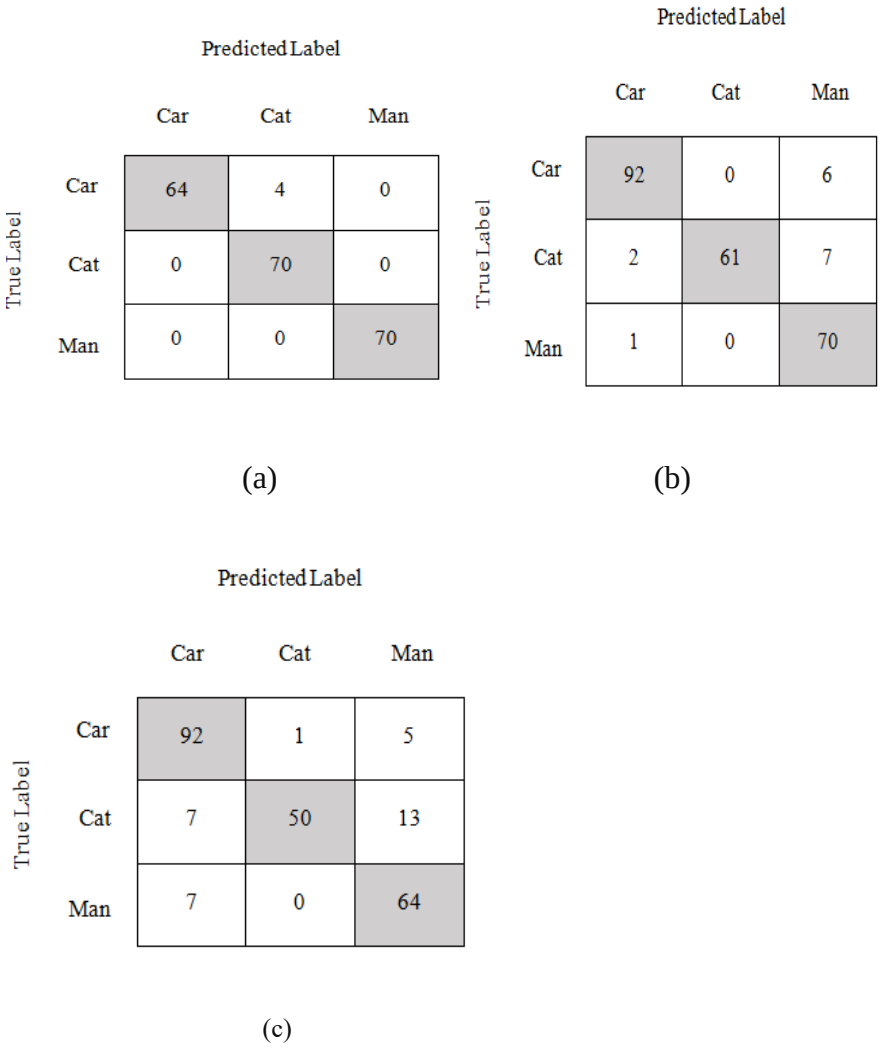


Figure 5.18: Confusion Matrix for FLIR (a) Inception v3 (b) MobileNet (c) DenseNet

Confusion matrix given in Figure 5.3 (a) depicts that for car class, 62 cars are predicted correctly while 4 cars are predicted as cat and 0 car is predicted as man, for cat class 70 cats are predicted as cats, 0 cat is predicted as car and 0 cat is predicted as man and For man class, 70 men are predicted as man, 0 men are predicted as cat and 0 man is predicted as car. Confusion matrix given in Figure 5.3 (b) depicts that for car class, 92 cars are predicted correctly while 0 cars are predicted as cat and 6 cars are predicted as man, for cat class 61 cats are predicted as cats, 2 cats are predicted as car and 7 cats are predicted as man and For man class, 70 men are predicted as man, 0 men are predicted as cat and 1 man is predicted as car. Confusion matrix given in Figure 5.3 (c) depicts that for car class, 92 cars are predicted

correctly while 1 car is predicted as cat and5 cars are predicted as man, for cat class 50 cats are predicted as cats, 7 cats are predicted as car and 13 cats are predicted as man and For man class, 64 men are predicted as man, 0 men are predicted as cat and 7 men are predicted as car.

The YOLOv4 Classification methodology was evaluated on Seek Thermal dataset as well as on FLIR. The dataset was split into 80:20 (≈5346, 1068) randomly as training and testing data respectively. There are three classes in the input data, cat class, car class and man class. The performance of the proposed model is depicted in the graphs below (figure 26) for both dataset. In the graphs the blue line is depicting the testing loss and the orange line is depicting the testing accuracy, the lines of the graph clearly shows that model is not under fit as well as not over fit. Testing accuracy is increasing with the number of steps. But as shown in graph with further increase in steps the accuracy is not increasing further.

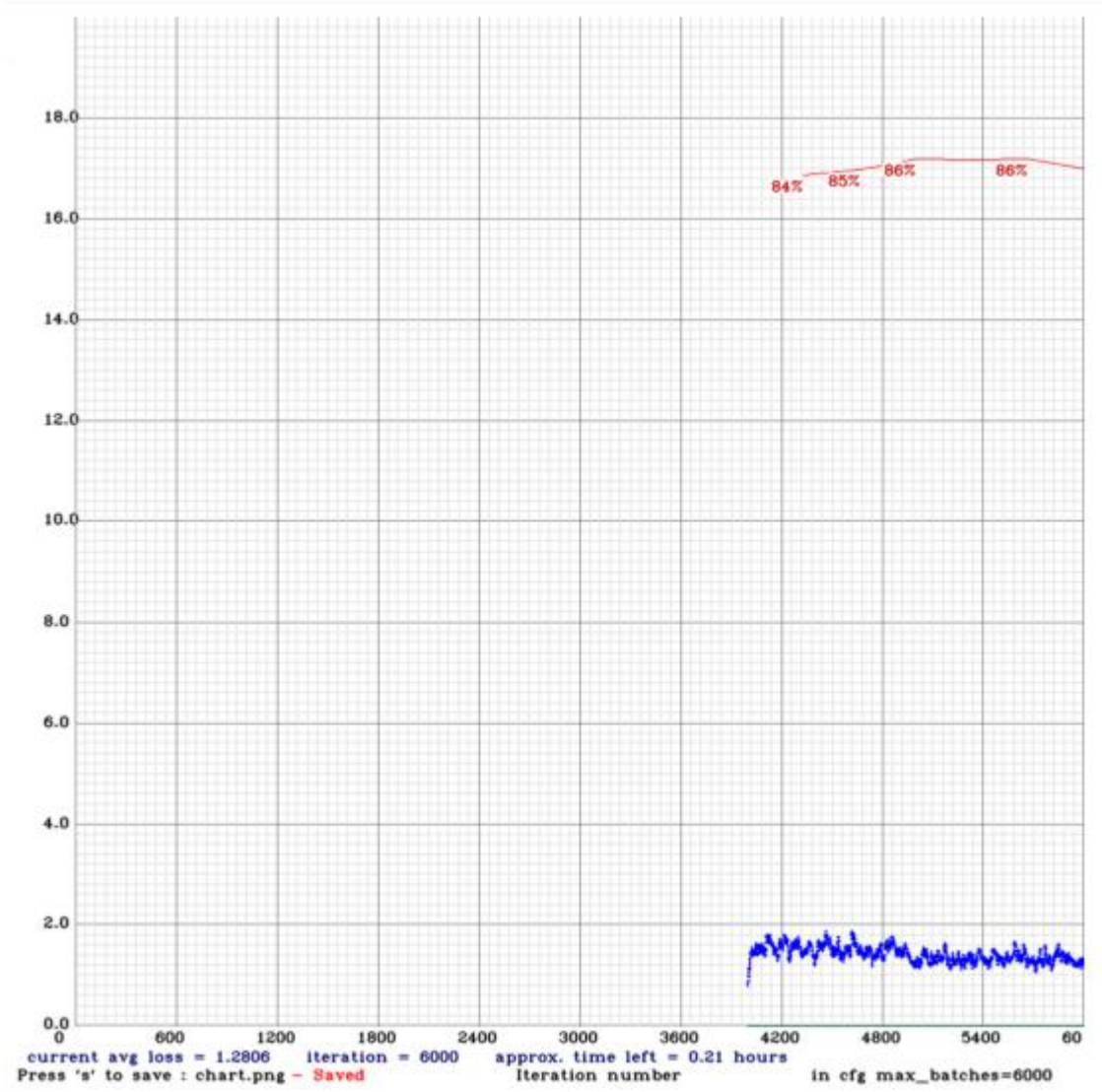


Figure 5.19: Accuracy plot YOLOv4 – Seek Thermal

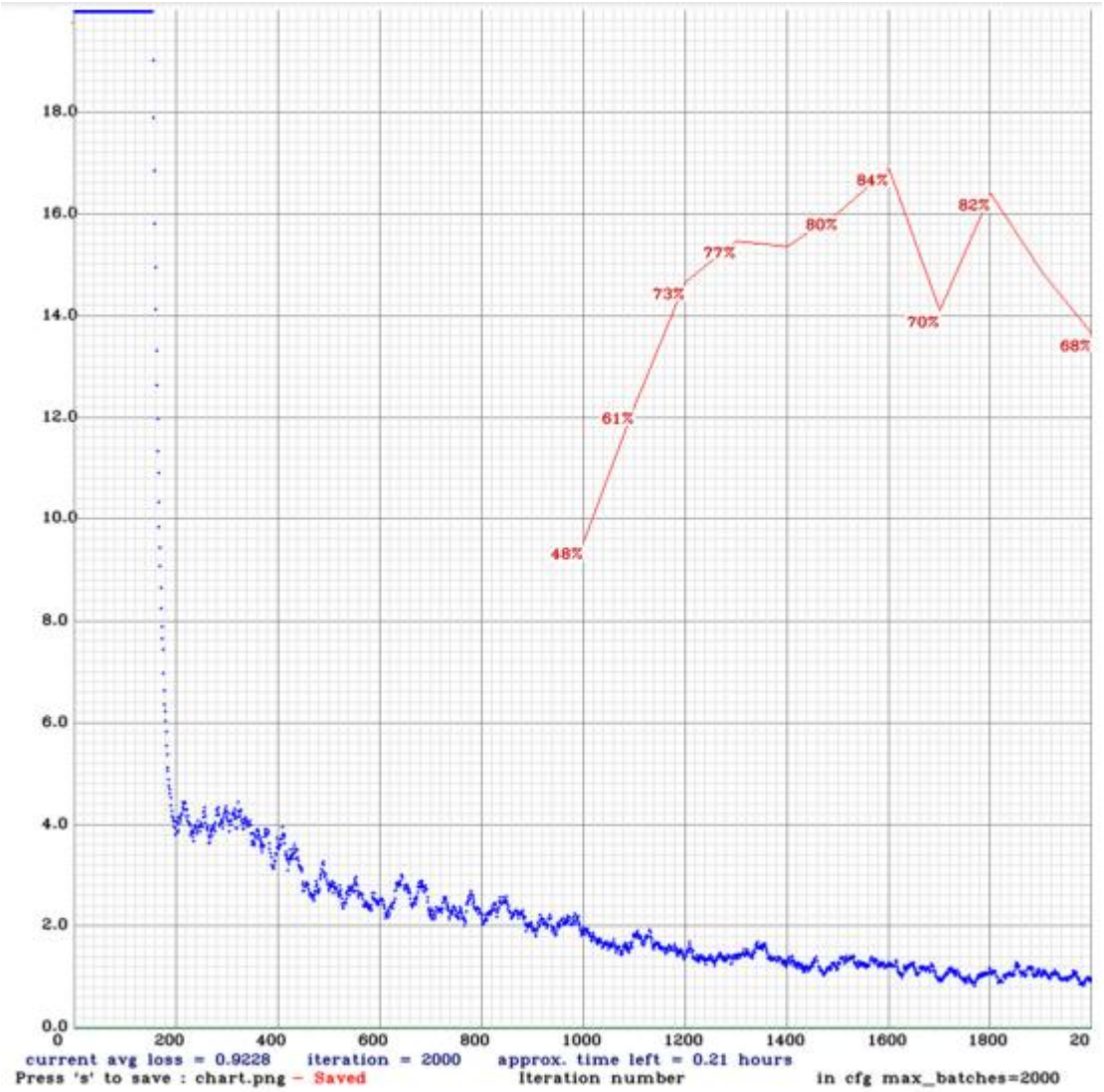


Figure 5.20: Accuracy plot YOLOv4 – FLIR

Prediction

True Label		TP	FP
	Car	322	32
	Cat	328	19
	Man	255	81

(a)

Prediction

True Label		TP	FP
	Car	61	40
	Cat	55	6
	Man	55	14

(b)

Figure 5.21: YOLOv4 TP and FP for (a) Seek Thermal (b) FLIR

Figure 5.6(a) shows that for car class true positives are 322, false positives are 32 and average precision is 91.76%, for cat class true positives are 328, false positives are 19 and average precision is 88.55% and for man class true positives are 255, false positives are 81 and average precision is 76.28%. Figure 5.6(b) shows that for car class true positives are 61, false positives are 40 and average precision is 78.08%, for cat class true positives are 55, false positives are 6 and average precision is 90.35% and for man class true positives are 55, false positives are 14 and average precision is 85.47%.

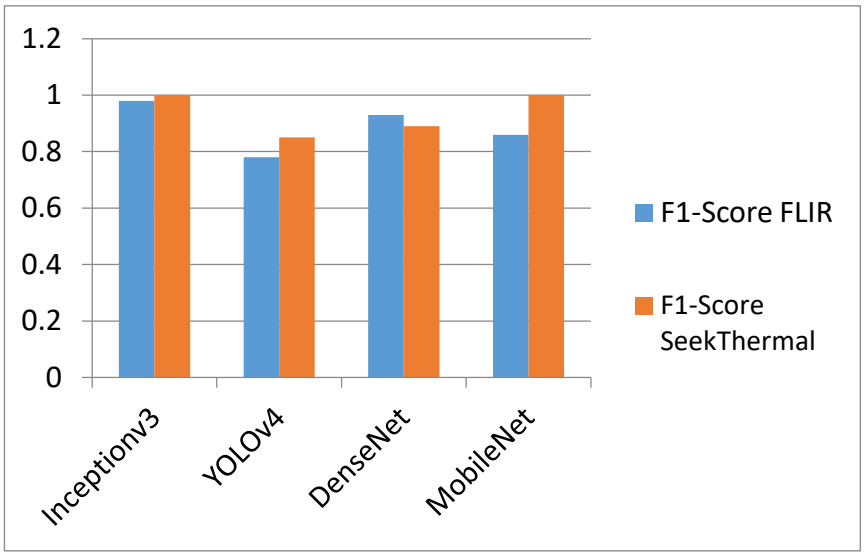


Figure 5.22: Comparison Graph for CNN models based on F1 - score

Inception v3 give highest results for the dataset obtained by using FLIR camera and by using Seek Thermal camera and DenseNet also give best results for the dataset obtained by using FLIR camera, MobileNet give best results for the dataset obtained by using Seek Thermal camera and YOLOv4 give average results for both Seek Thermal and FLIR. The proposed framework based on Inception v3 gives 0.98 F1-Score for FLIR and 1 for Seek Thermal. The proposed framework based on Inception v3 gives the highest accuracy and F1 – score for both Seek Thermal and FLIR that is 100 % and 98.91% respectively.

Comparison of the proposed framework with the customized neural network is shown in figure 5.7 is presented in table 5.5. This comparison shows that proposed framework give better results as compare to customized neural network , customized neural network needs more training and needs to be more deeper.

Database	Model	No. of Images	Training (80%)	Testing (20%)	Prec %	Recall %	WAP %	WAR %	ACC %	F1-Score	Miss Classified %
FLIR	Inceptionv3	1014	807	207	98.90	98.90	99	97	98.91	0.98	1.09
	Customized CNN				71.5	71.5	72	71.69	71.42	0.71	28.58
Seek Thermal	Inceptionv3	6414	5346	1068	100	100	100	100	100	0.100	0
	Customized CNN				95.59	95.59	95	94.7	95.59	0.96	4.41

Table 5.7: Comparison of Framework with a customized CNN

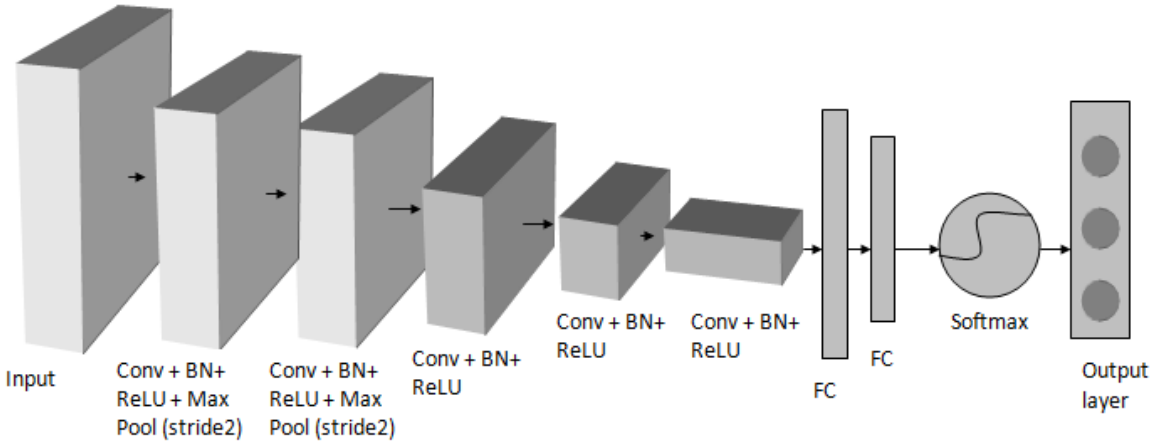
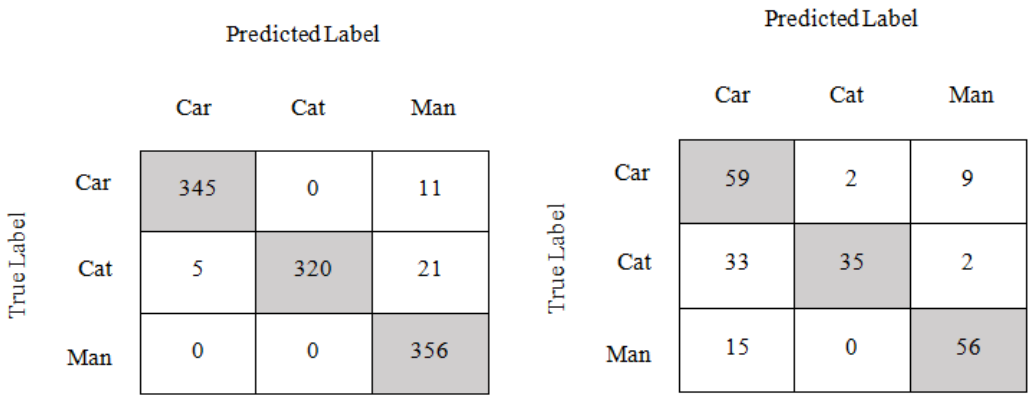


Figure 5.23 Customized CNN Architecture



(a) (b)

Figure 5.24: Confusion Matrix for Customized CNN (a) FLIR (b) Seek Thermal

Confusion matrix given in Figure 5.8 (a) depicts that for car class, 345 cars are predicted correctly while 0 cars are predicted as cat and 11 cars are predicted as man, for cat class 320 cats are predicted as cats, 5 cats are predicted as car and 21 cats are predicted as man and For man class, 356 men are predicted as man, 0 men are predicted as cat and 0 man is predicted as car. Confusion matrix given in Figure 5.3 (b) depicts that for car class, 59 cars are predicted correctly while 2 cars are predicted as cat and 9 cars are predicted as man, for cat class 35 cats are predicted as cats, 33 cats are predicted as car and 2 cats are predicted as man and For man class, 56 men are predicted as man, 15 men are predicted as car and 0 man is predicted as cat.

Comparison of proposed framework with a conventional method in the literature that is mostly used shown in table 5.6. It shows that CNN model Inception v3 gives best results as compared to HOG and SVM.

Database	Model	No. of Images	Training (80%)	Testing (20%)	Prec %	Recall %	WAP %	WAR %	ACC %	F1-Score	Miss Classified %
FLIR	Inceptionv3	1014	807	207	98.90	98.90	99	97	98.91	0.98	1.09
	HOG + SVM				72.51	72.51	79	73	72.51	0.72	27.49
Seek Thermal	Inceptionv3	6414	5346	1068	100	100	100	100	100	0.100	0
	HOG + SVM				86	86	72	72.48	86.5	0.86	13.5

Table 5.8: Comparison of Framework with a conventional method

Some conventional methods did thermal image classification but no work is done on the usage of deep learning for the classification of thermal images. Some international literature was found related to thermal image classification but no literature is related to usage of deep learning models for classification of objects like car, cat and man, so for comparison of framework dense net, mobile net and YOLOV4 models are trained and tested on the same dataset. Comparison with a conventional method and a customized CNN model is also presented.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion

A framework that is based on deep CNN (Convolutional Neural Network) is presented in this research for the classification of thermal images of multiple objects. The results show that a deep CNN (Convolutional Neural Network) is an asset for the classification of the thermal imaging. Some time is needed for the training of such models, but once a model gets trained in seconds it can give a comparable results. The framework is trained and tested on the thermal images that are captured by using FLIR and Seek Thermal. Transfer learning is applied for the training of model Inception v3 that is used for classification and transfer learning is used for training of models Yolov4, MobileNet and DenseNet that are used for comparison. A method in which a model that is trained on the one dataset is reused for another dataset is known as transfer learning. For training and testing of the model two datasets are used that include three classes' cat, car, and man. For FLIR dataset the highest accuracy achieved is 98.91% and for Seek thermal dataset highest accuracy achieved is 100%. A comparison of proposed framework with some other CNN models (DenseNet, MobileNet and YOLOv4), with customized CNN model and with conventional models is also presented. The results of proposed framework and comparison with other models prove that proposed framework is effective for the classification of thermal images.

6.2 Contribution

- A framework is presented for the classification of thermal images of multiple objects.
- A good dataset of thermal images for object classification is provided to the research community

6.3 Future Work

In conventional algorithms of machine learning, to train a classifier different features are manually selected. Deep learning models extract the features by themselves no need of extracting features set according to the problem. In future the dataset can be extended by adding more classes related to daily life objects. A more deep customized CNN model can be designed to optimize resource utilization along with good accuracy.

REFERENCES

[1] T Z. Jia, Z. Liu, C. Vong and M. Pecht, "A Rotating Machinery Fault Diagnosis Method Based on Feature Learning of Thermal Images," in *IEEE Access*, vol. 7, pp. 12348-12359, 2019.

[2] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park and K. R. Park, "Action Recognition From Thermal Videos," in *IEEE Access*, vol. 7, pp. 103893-103917, 2019.

[3] European Space Agency. [Online], https://sci.esa.int/web/education/-/48986-blackbody-radiation, accessed on Dec, 2019.

[4] E. Jackson and L. Chermak, "PUGTIFs: Passively User-Generated Thermal Invariant Features," in *IEEE Access*, vol. 7, pp. 109566-109576, 2019.

[5] TechTarget. [Online], https://whatis.techtarget.com/definition/thermal-imaging, accessed on Dec, 2019.

[6] Towards Data Science. [Online], https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, accessed on Dec, 2019.

[7] S. Menon, S. J., A. S.K., A. P. Nair and S. S., "Driver Face Recognition and Sober Drunk Classification using Thermal Images," *2019 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 2019, pp. 0400-0404.

[8] Medium. [Online], https://medium.com/@zurister/depth-wise-convolution-and-depth-wise-separable-convolution-37346565d4ec, accessed on Jan, 2020.

[9] Machine Think. [Online], https://machinethink.net/blog/mobilenet-v2/, accessed on Jan, 2020.

[10] M. Haider, A. Doegar and R. K. Verma, "Fault Identification in Electrical Equipment using Thermal Image Processing," *2018 International Conference on Computing, Power and*

Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 853-858.

[11] G. Lu, H. Yu and C. Yuan, "Getting Rid of Night: Thermal Image Classification Based on Feature Fusion," *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, 2018, pp. 2827-2832.

[12] C. K. Kyal, H. Poddar and M. Reza, "Detection of Human Face by Thermal Infrared Camera Using MPI model and Feature Extraction Method," *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1-5.

[13] Towards Data Science. [Online], https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728, accessed on Jan 2020.

[14] Deepai. [Online], https://deepai.org/machine-learning-glossary-and-terms/batch-normalization, accessed on Jan 2019.

[15] E. Bartuzi, K. Roszczewska, A. Czajka and A. Pacut, "Unconstrained biometric recognition based on thermal hand images," *2018 International Workshop on Biometrics and Forensics (IWBF)*, Sassari, 2018, pp. 1-8.

[16] S. Wang, B. Pan, H. Chen and Q. Ji, "Thermal Augmented Expression Recognition," in *IEEE Transactions on Cybernetics*, vol. 48, no. 7, pp. 2203-2214, July 2018.

[17] V. Ghenescu, E. Barnoviciu, S. Carata, M. Ghenescu, R. Mihaescu and M. Chindea, "Object Recognition on Long Range Thermal Image Using State of the Art DNN," *2018 Conference Grid, Cloud & High Performance Computing in Science (ROLCG)*, Cluj-Napoca, 2018, pp. 1-4.

[18] Jian-Feng Shi, S. Ulrich and S. Ruel, "A comparison of feature descriptors using monocular thermal camera images," *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*, Nagoya, 2017, pp. 225-228.

[19] D. Zhou and Jingzhou Wang, "Identification of deer in thermal images to avoid deer-vehicle crashes," *Proceedings of 2011 International Conference on Electronics and Optoelectronics*, Dalian, 2011, pp. V3-342-V3-345.

[20] D. Kim and D. Kwon, "Pedestrian detection and tracking in thermal images using shape

features," *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Goyang, 2016, pp. 22-25.

[21] J. Hossen, E. L. Jacobs and F. K. Chowdhury, "Activity recognition in thermal infrared video," *SoutheastCon 2015*, Fort Lauderdale, FL, 2015, pp. 1-2.

[22] Yan Zhang, Ting Zhao, Jian Gu and Shengyang Yu, "Accurate moving object detection in thermal imagery," *2011 IEEE International Conference on Computer Science and Automation Engineering*, Shanghai, 2011, pp. 282-286.

[23] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[24] Kaggle. [Online], https://www.kaggle.com/hexietufts/easy-to-use-keras-imagedatagenerator, accessed on Jan, 2019.

[25] Hacker Earth. [Online], https://www.hackerearth.com/practice/machine-learning/transfer-learning/transfer-learning-intro/tutorial/, accessed on Mar, 2020.

[26] Towards DataScience. [Online], https://towardsdatascience.com/yolo-v4-optimal-speed-accuracy-for-object-detection-79896ed47b50, accessed on Sep, 2020

[27] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

[28] Dargan, Shaveta, et al. "A survey of deep learning and its applications: A new paradigm to machine learning." *Archives of Computational Methods in Engineering* (2019): 1-22.

[29] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).

[30] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[31] Wang, Wei, et al. "A novel image classification approach via dense-MobileNet models." Mobile Information Systems 2020 (2020).

[32] Machine learning Mastery. [Online] , https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/, accessed on Mar, 2021

[33] Grainger.[Online], https://www.grainger.com/content/qt-thermal-imaging-applications-uses-features-345, accessed on Mar, 2021

[34] Science Direct. [Online], https://www.sciencedirect.com/topics/earth-and-planetary-sciences/thermal-imaging, accessed on Mar, 2021

[35] Crisp.[Online], https://crisp.nus.edu.sg/~research/tutorial/em.htm, accessed on Mar, 2021.

[36] Researchgate. [Online], https://www.researchgate.net/publication/327077169_Review_of_Biomedical_Applications_of_Contactless_Imaging_of_Neonates_using_Infrared_Thermography_and_Beyond/figures?lo=1, accessed on Mar, 2021.

[37] Thermal imaging camera reviews.[Online], https://thermalimagingcamerareviews.com/thermal-camera-manufacturers/, accessed on Mar, 2021.

[38] If Sec global.[Online], https://directory.ifsecglobal.com/thermal-imaging-code004833.html, accessed on Mar,2021

[39] The architects guide.[Online], https://www.thearchitectsguide.com/articles/best-thermal-imaging-camera, accessed on Mar, 2021

[40] Thermogears.[Online], https://thermogears.com/guide-choose-best-thermal-imaging-cameras/, accessed on Mar, 2021

[41] linkedin. [Online], https://www.linkedin.com/company/seek-thermal, accessed on Mar, 2021

[42] Christian Szegedy et al., "Going Deeper with Convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

[43] Markets and markets.[Online], https://www.marketsandmarkets.com/Market-Reports/thermal-imaging-market-1300.html, accessed on Mar, 2021

[44] FLIR.[Online], https://www.flir.com/oem/adas/adas-dataset-form/, accessed on Mar, 2021

[45] Imagine.[Online], https://imagine.gsfc.nasa.gov/science/toolbox/emspectrum1.html,accessed on Mar, 2021.

[46] FLIR.[Online], https://www.flir.eu/discover/rd-science/how-do-thermal-cameras-work/, accessed on Mar, 2021.