VISUAL SENTIMENT ANALYSIS

A PROJECT REPORT

Submitted By

BHARATH KUMAR A M R. 185001034

MRUDUL PRABANDH P. 185001095

SANTHOSH R. 185001135

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

Rajiv Gandhi Salai (OMR), Kalavakkam - 603110

May 2022

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this project report titled **"VISUAL SENTIMENT ANALYSIS"** is the *bonafide* work of "**BHARATH KUMAR A M R (185001034)**, **MRUDUL PRABANDH P (185001095)**, and **SANTHOSH R (185001135)**" who carried out the project work under my supervision.
Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Dr. T.T. Mirnalinee
Head of the Department
Professor,
Department of CSE,
SSN College of Engineering,
Kalavakkam - 603 110

Dr. T.T. Mirnalinee
Supervisor
Professor,
Department of CSE,
SSN College of Engineering,
Kalavakkam - 603 110

Place:
Date:

Submitted for the examination held on............

Internal Examiner

External Examiner

ACKNOWLEDGEMENTS

I thank GOD, the almighty for giving me strength and knowledge to do this project.

I would like to thank and deep sense of gratitude to my guide **Dr. T.T. MIRNALINEE**, Professor, Department of Computer Science and Engineering, for his valuable advice and suggestions as well as his continued guidance, patience and support that helped me to shape and refine my work.

My sincere thanks to **Dr. T.T. MIRNALINEE**, Professor and Head of the Department of Computer Science and Engineering, for her words of advice and encouragement and I would like to thank our project Coordinator **Dr.B. BHARATHI**, Associate Professor, Department of Computer Science and Engineering for her valuable suggestions throughout this project.

I express my deep respect to the founder **Dr. SHIV NADAR**, Chairman, SSN Institutions. I also express my appreciation to our **Dr. V. E. ANNAMALAI**, Principal, for all the help he has rendered during this course of study.

I would like to extend my sincere thanks to all the teaching and non-teaching staffs of our department who have contributed directly and indirectly during the course of my project work. Finally, I would like to thank my parents and friends for their patience, cooperation and moral support throughout my life.

BHARATH KUMAR A M R. MRUDUL PRABANDH P. SANTHOSH R.

ABSTRACT

Visual Sentiment Analysis aims to understand how an image evokes the emotions of the people. Sentiment analysis has been mostly done on text and nowadays people started conveying their emotions through images and videos which led to the need of sentiment analysis on images. The challenges related to sentiment analysis on images are to capture the emotions of all objects in the image, the data for those analysis are not enough so the collection of data becomes a challenge and to how to implement the project using deep leaning methods. Our Objective is to do sentiment analysis on the natural disaster images where mostly all images are negative but few images convey positive sentiments. To classify those positive images from negative images by the analysis of the image. An image also conveys more than one sentiments so we are also doing in depth analysis of image to show all the sentiments related to the image. The best suit models for this projects are VGG19, efficient net and vision transformer. These models are used for single label and multi label classification of the natural disaster images. Finally combining all the three models to give a ensemble output. Our model achieved 0.85 weighted average F1 score for single label classification and 0.71 for multi label classification. This project also includes the new challenges and also the insights from the study of the results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As of late, the public accessibility of data on the web has provoked the review and improvement of Sentiment Analysis, which examinations a lot of client information to deduce responses to subjects, suppositions, and patterns to more readily comprehend the state of mind of clients who make and offer data on the web. The objective of opinion investigation is to remove individuals' mentalities on a point or the writer's expected close to home impact on the perusers.

The undertakings of this exploration field are trying as well as extremely valuable. Feeling examination tracks down a few viable applications, since sentiments impact numerous human choices either in business and social exercises. As case, organizations are keen on observing individuals suppositions toward their items or administrations, as well as clients depend on criticisms of different clients to assess an item before they buy it. With the development of online entertainment (i.e., audits, discussions, sites and informal communities), people and associations are progressively involving popular suppositions for their independent direction.

The fundamental errand in Sentiment Analysis is the extremity characterization of an information message (e.g., taken from a review,a remark or a social post) concerning positive, negative or impartial extremity. This investigation can be performed at record, sentence or component level. The techniques for this area are valuable to catch popular assessment on items, administrations, promoting, political inclinations and get-togethers. For instance the investigation of the action of Twitter's clients can assist with anticipating the ubiquity of gatherings or

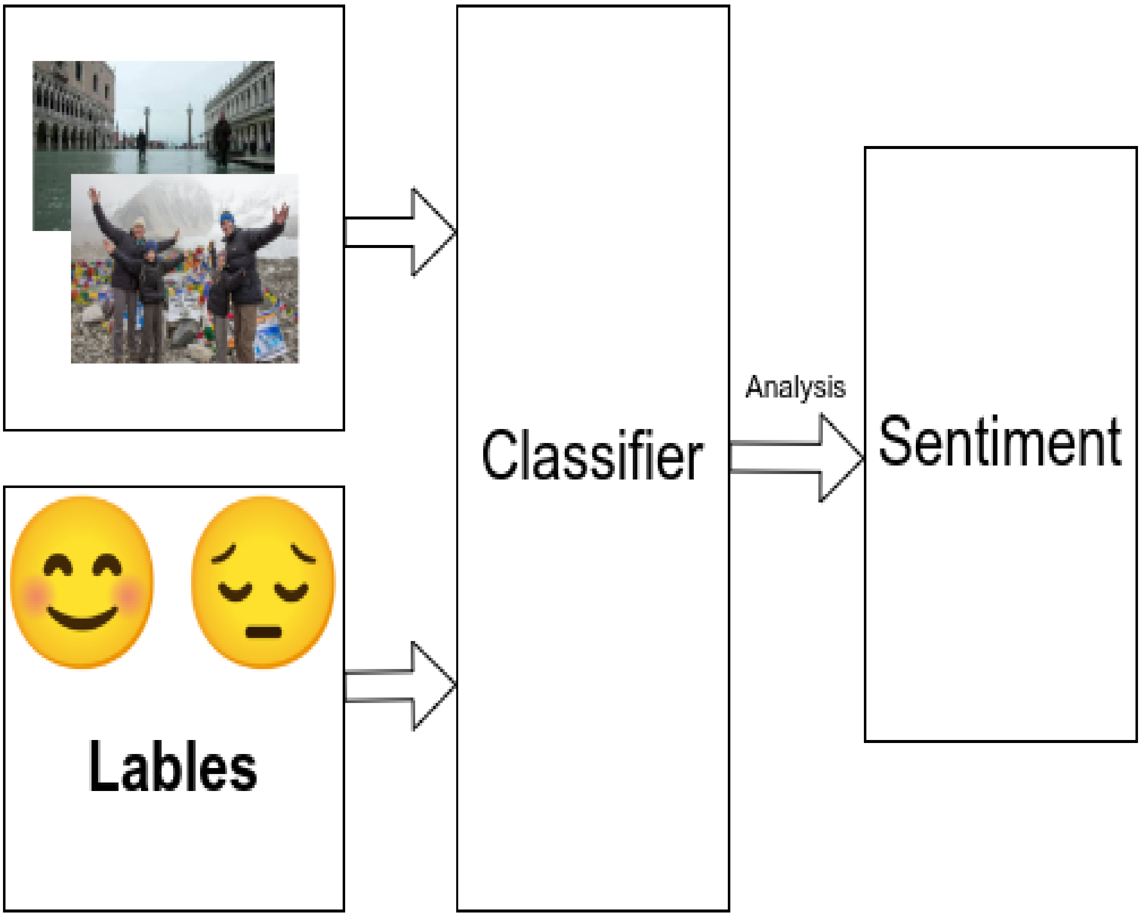


FIGURE 1.1: Diagram of a generic visual sentiment analysis framework

alliances. The accomplished outcomes in Sentiment Analysis inside miniature writing for a blog have shown that Twitter posts sensibly mirror the political scene. By and large, Sentiment Analysis procedures have been created for the investigation of message ,though restricted endeavors have been utilized to remove (i.e., deduce) opinions from visual items (e.g.,images and recordings).

Despite the fact that the logical exploration has previously accomplished eminent outcomes in the field of literary Sentiment Analysis in various settings (e.g., informal organization posts examination, item surveys, political inclinations, and so on) the errand to comprehend the state of mind from a message has a few hardships given by the intrinsic vagueness of the different dialects (e.g., unexpected sentences), social elements, etymological subtleties and the trouble of sum up any message investigation answer for various language vocabularies. The various arrangements in the field of message Sentiment Analysis have not yet accomplished a degree of dependability sufficient to be carried out without encasing the connected setting. For instance, notwithstanding the presence of regular language handling devices for the English language, similar apparatuses can't be utilized straightforwardly to dissect text written in different dialects. Despite the fact that NLP (Natural Language Processing) and data recovery analysts proposed a few ways to deal with address the issue of Sentiment Analysis, the online entertainment setting offers a few extra difficulties. Close to the immense measures of accessible information, regularly the printed interchanges on interpersonal organizations comprise of short and casual messages.

Besides, individuals will generally utilize likewise pictures and recordings, notwithstanding the printed messages, to communicate their encounters through the most well-known social stages. The data contained in such visual items are not just connected with semantic items like articles or activities about the procured picture, yet in addition signs about effect and feeling conveyed by the portrayed scene. Such data is consequently valuable to grasp the close to home effect (i.e., the evoked feeling) past the semantic.

Thus pictures and recordings have become one of the most famous media by which individuals express their feelings and offer their encounters in the interpersonal organizations, which play expected a urgent part in gathering information about individuals' viewpoints and sentiments. The pictures partook in online entertainment stages reflect visual parts of clients' everyday exercises and interests. Such developing client created pictures address a new and strong wellspring of data valuable to investigate clients' inclinations. Separating the feelings intrinsically hidden pictures apparent by the watchers could elevate new ways to deal with a few application fields, for example, brand discernment, evaluation of consumer loyalty, publicizing, media logical, and so forth.

The dissemination of individual cell phones continually associated with Internet administrations (e.g., cell phones) and the development of online entertainment stages presented another correspondence worldview by which individuals that share sight and sound information. In this unique circumstance, transferring pictures to a web-based entertainment stage is the new way by which individuals share their perspectives and encounters. This gives serious areas of strength for a to explore on this field, and offers many testing research issues.

The proposed project aims to introduce this emerging research field. In this project we are going to analyze the sentiments related to natural disaster images. The focus of this project is images related to different natural disasters around the world, able to perform a meaningful analysis on this data could be of great societal importance. The challenges related to this topic are considering both objects,scenario at the same time and giving importance to one which is more being appropriate for the image and the amount of data available to do these types analysis are low. So the collection of data for these tasks are one of the important challenge. An image can express more than one sentiments at a same time, our main challenge is to analyse all the sentiments related to the images.

In recent years, deep learning convolutional neural networks have produced better results for visual based projects. In our projects we will be using two CNN networks VGG19 and EfficientNet. In addition to that we will also be using Vision transformer model because they outperforms CNN when it comes to facial feature recognition. In this paper we will be using mediaeval visual sentiment dataset with single label and multi label annotations for each images. Result Analysis reports the insights as outcome of the proposed study. Then, with comparison from the existing works, concludes the project by providing a summary of the previous sections and by suggesting directions for future works.

CHAPTER 2

LITERATURE SURVEY

In [6] Deep learning has arisen as a strong AI strategy that learns numerous layers of portrayals or elements of the information and produces best in class forecast results. In this the creator delineates the significance of public exploring for instance. Allow us to consider a model as though one needs to purchase a customer item, one is not generally restricted to approaching one's loved ones for conclusions since there are numerous client surveys and conversations about the item in open discussions on the Web. For an association, it might at this point not be important to direct overviews, assessments of public sentiment, and center gatherings to assemble popular feelings since there is an overflow of such data freely accessible.

Lately, they have seen that stubborn postings in web-based entertainment have reshaped organizations, and influence public feelings and feelings, which have significantly affected on our social and political frameworks. In any case, finding and observing assessment destinations on the Web and refining the data contained in them stays an impressive errand due to the multiplication of different locales. Each webpage ordinarily contains an enormous volume of assessment text that isn't generally handily unraveled in lengthy sites and discussion postings. The typical human peruser will experience issues recognizing significant locales and extricating and summing up the sentiments in them. Computerized opinion investigation frameworks are hence required.

In [4] Sentiment examination alludes to the administration of feelings, conclusions, and emotional message. The interest of opinion examination is raised because of the necessity of breaking down and organizing stowed away data, extricated from online entertainment as unstructured information. The feeling examination is being executed through profound learning methods. Profound learning networks are superior to SVMs and typical brain networks since they have more secret layers when contrasted with typical brain networks that have a couple of stowed away layers. Profound learning networks are proficient to give preparing in both regulated/solo ways. Profound Learning networks do programmed highlight extraction and doesn't include human mediation accordingly it can save time since include designing isn't required.

In [10] Visual opinion examination by means of profound various bunched occasion learning organization (DMCILN) is developed for visual feeling investigation. The info picture is changed over into a sack of cases through a visual occasion age module, which is made out of a pre-prepared CNN and two transformation layers. Upgrades required: As DMCILN is utilized, pack of occurrences consume an enormous space contrasted with other datasets.

Machajdik et al.[5] did a concentrate on utilizing separated highlights in view of brain research and workmanship hypothesis to group the profound reaction of a given picture. The elements were gathered by variety, texture,composition, and content, and afterward characterized by an innocent Bayes based classifier. Upgrades required : Although the work accomplished great outcomes for the time, the removed elements struggle with catching the complicated connection between human inclination and the substance of a picture.

In [8] Visual opinion examination is an extremely difficult errand and spotlights on picture feeling examination utilizing convolutional brain organizations. It utilized the idea of averaging for producing opinion probabilities and it might actually act as a basic methodology. Upgrades required: It doesn't deliver an advanced arrangement where in other DL fields are more improved answers for creating more precise outcomes.

In [11] A visual opinion examination system was presented utilizing an exceptionally profound convolutional brain network with Transfer discovering that beats past utilized standard dataset. This model can be utilized to break down limited scope information mixed media content for figuring out client input, publicizing, and prescient displaying. Upgrades required: It showed the way that CNN can perform well with a more modest dataset however with a huge scope dataset it would be truly hard to utilize Transfer Learning approach.

In [9] the issue of visual opinion examination in view of convolutional brain organizations, where the feelings are anticipated.. The recognition branch is intended to naturally take advantage of the feeling map, which can give the limited data of the emotional pictures. Then, at that point, the grouping branch utilizing both comprehensive and limited portrayals can anticipate the feelings. Exploratory outcomes show the viability of our technique.

In [7] they present the difficult issue of visual feeling examination. Because of the inaccessibility of a huge scope very much named informational collection, little exploration work has been distributed on concentrating on the effect of Convolutional Neural Networks on visual feeling investigation. In this work, they are presenting such an informational index and plan to deliver the informational collection to the exploration local area to advance the examination on visual feeling investigation with the profound learning and other learning systems. In the interim, they likewise assess the profound visual highlights separated from distinctively prepared brain network models.

Our trial results recommend that profound convolutional brain network highlights beat the condition of-theart hand-tuned highlights for visual feeling examination. Likewise, adjusted brain network on feeling related informational collections can additionally work on the exhibition of profound brain nework. By the by, the outcomes got in this work are just a beginning for the examination on utilizing profound learning or other learning structures for visual feeling investigation.

In [3] utilized the current pre-prepared move learning models, including the VGG-19, DenseNet-121, and ResNet50V2 models, for picture opinion expectation on the Crowdflower feeling extremity dataset. It presented an exceptional methodology that utilizations calibrated move learning models to deal with the issues of picture feeling examination. To relieve overfitting, they utilized extra layers, like dropout and weight regularization (L1 and L2 regularization). By playing out a matrix search across a few upsides of the regularization parameter,they had the option to track down the worth that gives the model its greatest exactness.

In [1] they have said about how to use vision transformer on visuals by dividing it into patches and also vision transformer outperforms all the CNN networks in terms of detecting facial expressing as well as the classification of visuals with less computations for training.

As far from the literature survey from the research papers, we have concluded that VGG-19, Efficient Net has been effective for visual image sentiment analysis. VGG-19 provides significant difference in accuracy when compared to the other models in VGG. Efficient Net provides high accuracy from less components, thus providing the efficient conditions for sentiment analysis. Vision Transformer as far has been effective for various models capturing the essentials in an image with detailed structures leading to an higher accuracy when compared to other transformer. We are further combining these three efficient models to ensemble (combining VGG, Efficient Net, Vision Transformer) to provide a better accuracy from ensembling.

CHAPTER 3

ARCHITECTURAL DESIGN FOR PROPOSED SYSTEM

In this section, we present our original visual opinion examination structure. Our proposed network structure is roused by VGG19, EfficientNet and Vision Transformer models

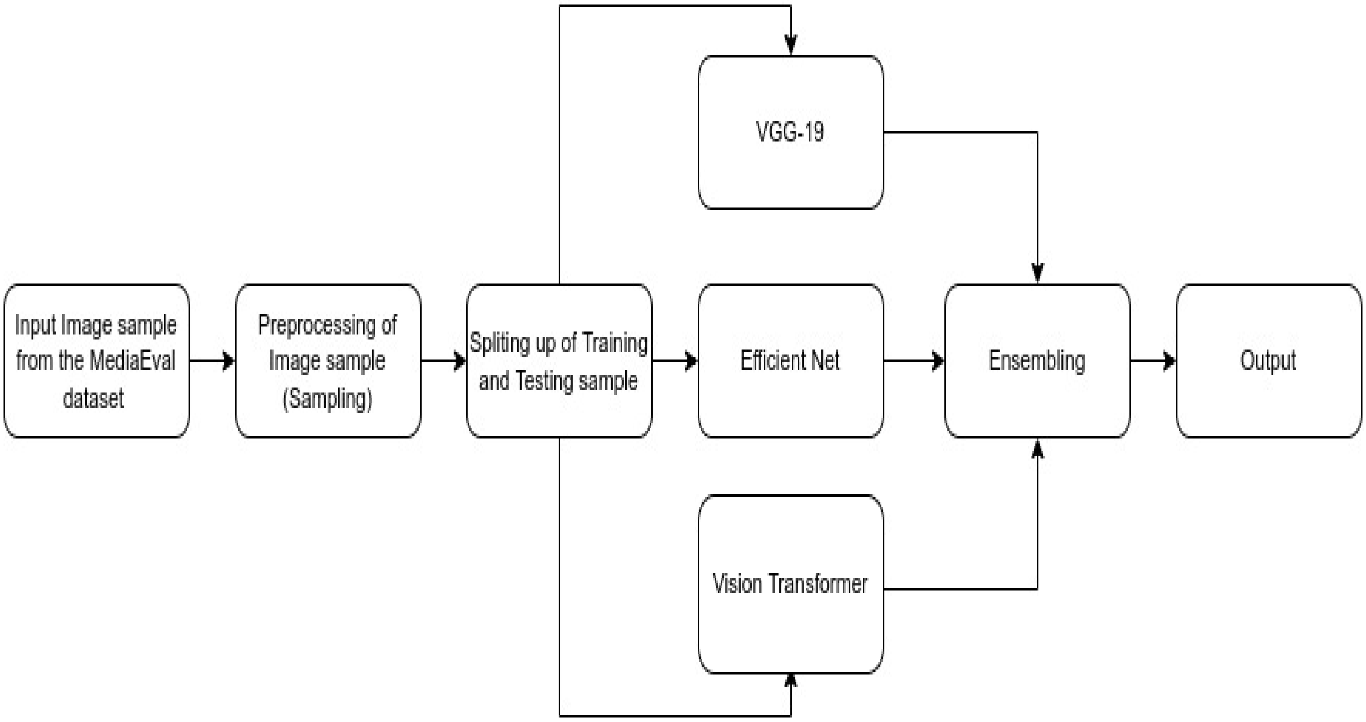

Input Image sample from the MediaEval dataset
Preprocessing of Image sample (Sampling)
Spliting up of Training and Testing sample
VGG-19
Efficient Net
Vision Transformer
Ensembling
Output

FIGURE 3.1: Architectural Design

3.1 OVERVIEW

The overview of this project is to develop a classifier ensemble model by combining VGG19, EfficientNet and Vision Transformer. This classifier model is trained with natural disaster images for single label sentiments as well as multi label sentiments. The output of this classifier model is to predict the sentiments evoked by the given image.

3.2 PREPROCESSING

By preprocessing we can defeat twists and upgrade a few highlights which are important for the specific application we are working for. We have utilized different preprocessing strategies for various models.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. The standard pre-processing were done to clean the raw data and make it suitable for a building and training the models by changing the input size of the image.

On this project, the preprocessing we have done are sampling using SMOTE, Early stopping handling and resizing of the input. These are further explained in the Implementation part.

3.3 VGG 19

Based on the literature survey [1] we found that convolutional neural network are better for image classification. In that VGGnet have better classification of images related to our topic based on the study from [2]. The reason why VGG19 was chosen is because it has 19 convolutional layers to increase in depth analysis of an image and it can also increase the model performance. In VGG19 the analysis of hidden layer is better when compared to other models. Since we have to analyse all the objects in the image to make a better classification related our topic so we are using VGG19. VGG hidden layers don't have local response normalisation. And also it increases memory consumption. The model diagram for VGG19 is given in Figure 3.2.

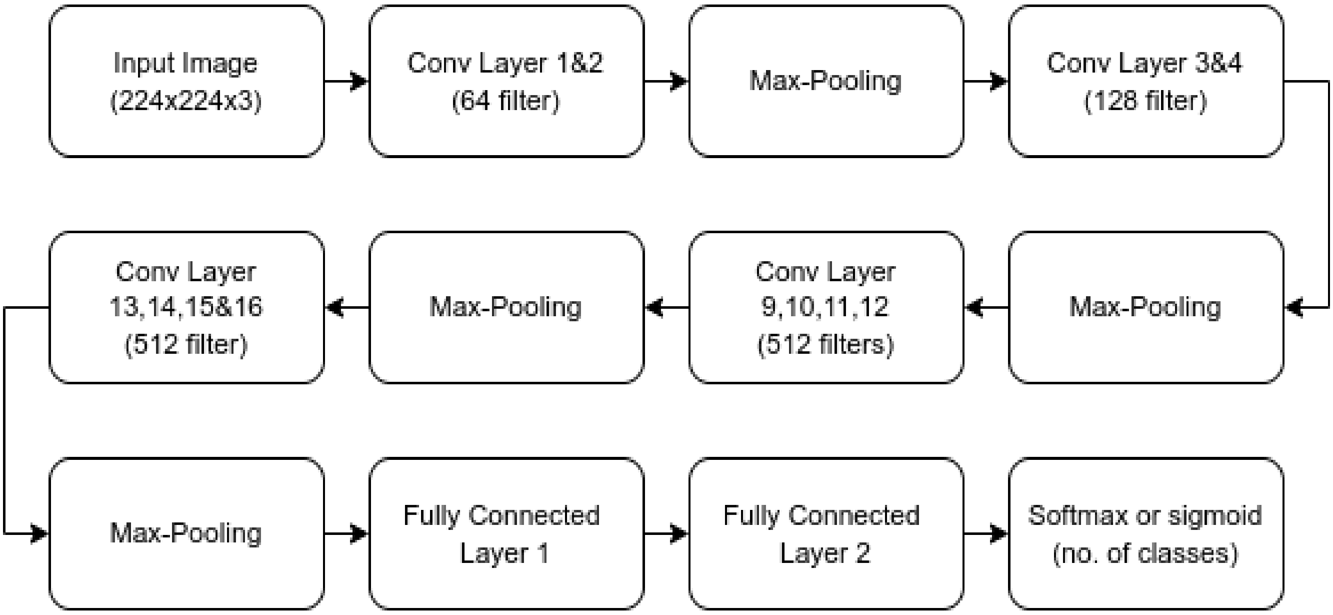


FIGURE 3.2: VGG-19 Model

Exception of VGG19 is that it takes more parameters which eventually takes more time to train a huge dataset.

3.4 EFFICIENTNET

As far the study from the paper[2]. We also found that EfficientNet also gives better classification with use less parameters and in less time. Not all models gives better accuracy by increase the layer. so our alternate approach CNN based model is EfficientNet. Compound coefficient to scale up models is the technique used by EfficientNet(given in Figure 3.3.) . These compound scaling method is based on the idea of balancing dimensions of width,depth and resolution(Figure 3.4.).The compound scaling method is to increase the size of the model. EfficientNet has many layers and FLOPS(Floating Point Operations Per Second) [6].

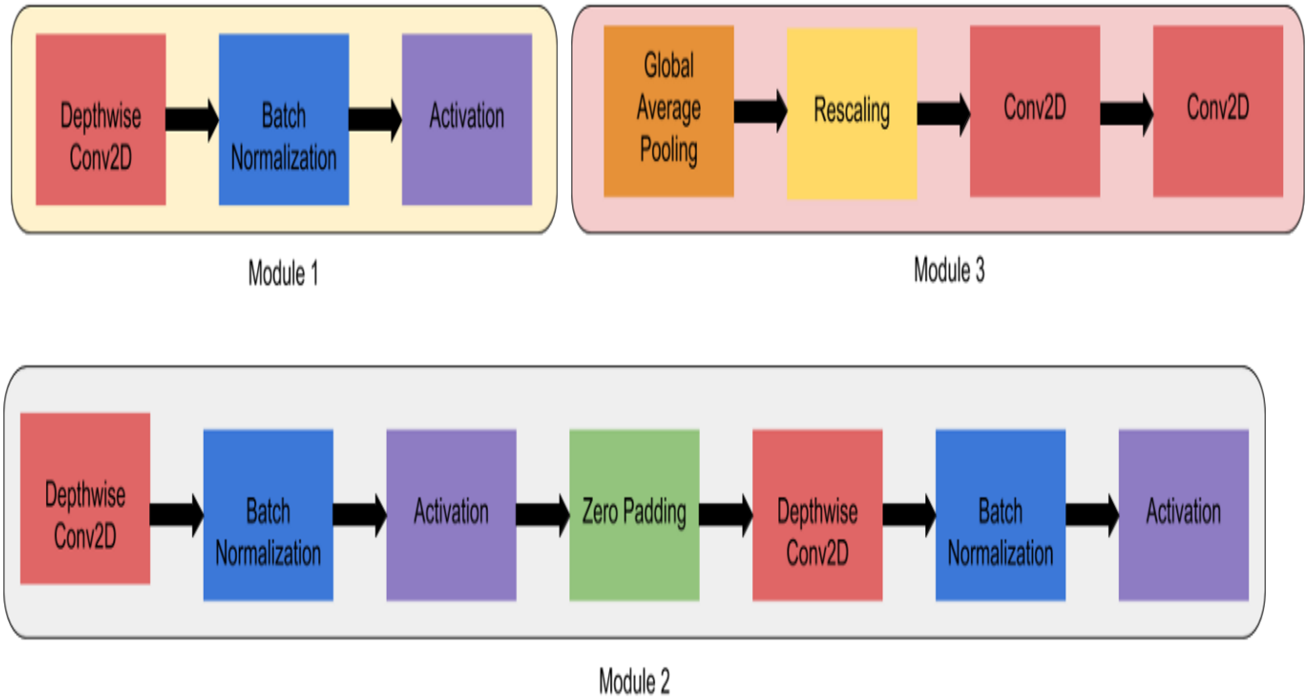


FIGURE 3.3: SCALING MODULES

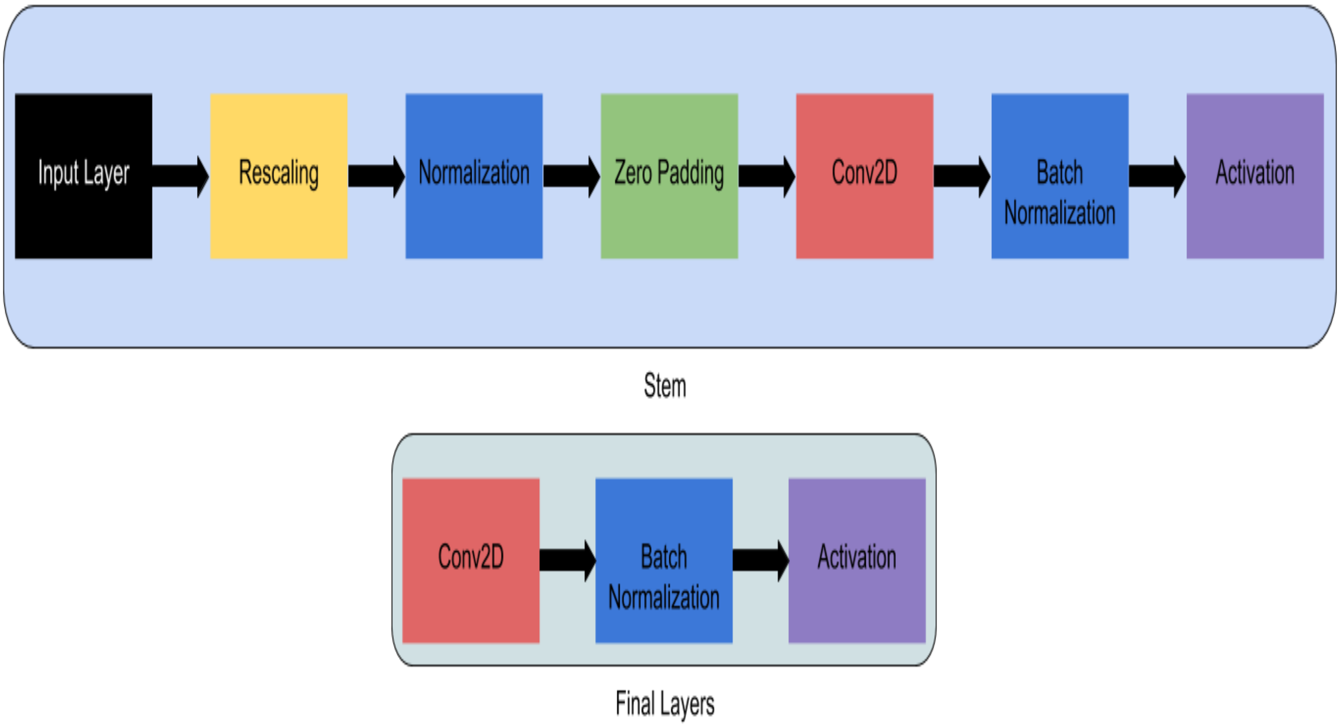


FIGURE 3.4: STEM AND FINAL LAYERS

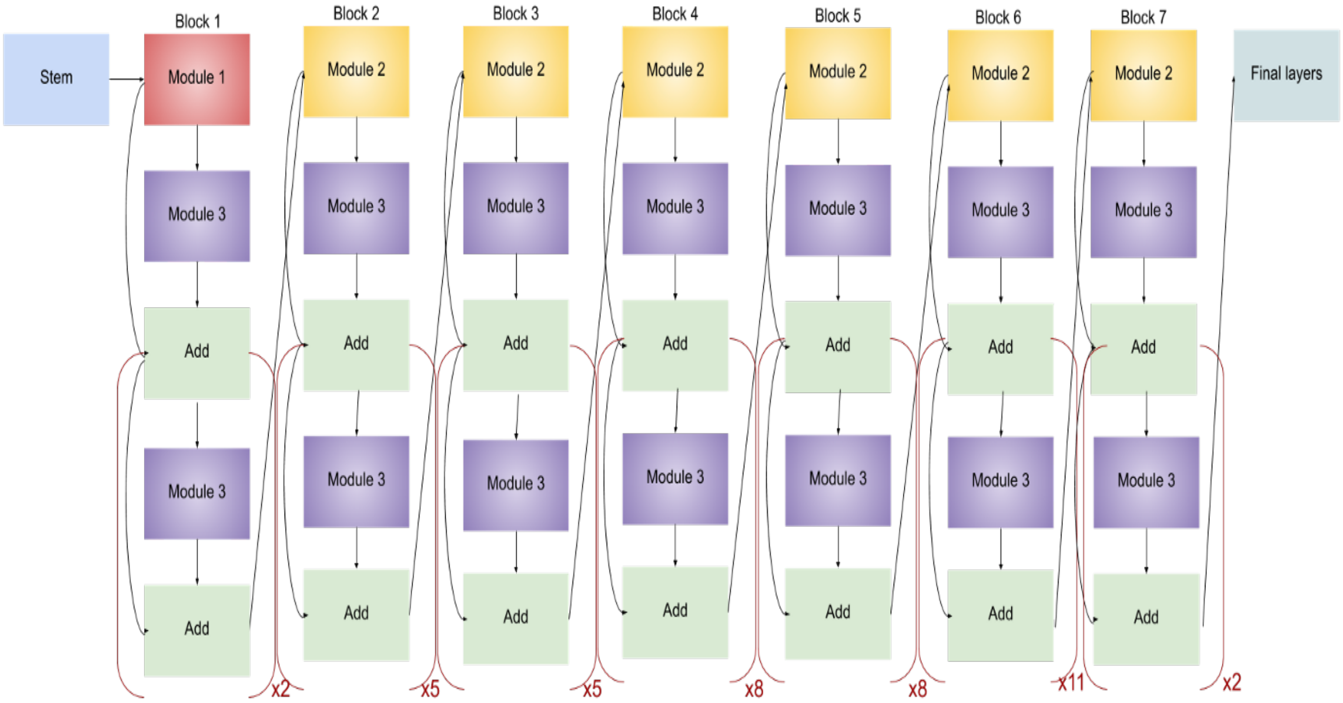


FIGURE 3.5: EfficientNet Architecture

3.5 VISION TRANSFORMER

We have used Vision Transformer an NLP based model which attains better results than CNN based model with less computational resources to train the model.The Vision Transformer, or ViT, is a model for image classification that employs a Transformer-like architecture over patches of the image. An image is split into patches the number of patch will be calculated based on the image size and patch size, each of the patches will be encoded, position encoding are also added, and the resulting sequence of vectors is fed to a standard Transformer encoder. In order to perform classification, the self attention head are used. Vision Transformer performs classification slightly better when compared to convolutional Networks when provided with huge data. Since CNN fail to encode relative facial information like vision Transformer. The vision transformer uses multi head self attention[7]. The overall working model diagram for vision transformer is given in Figure 3.6.

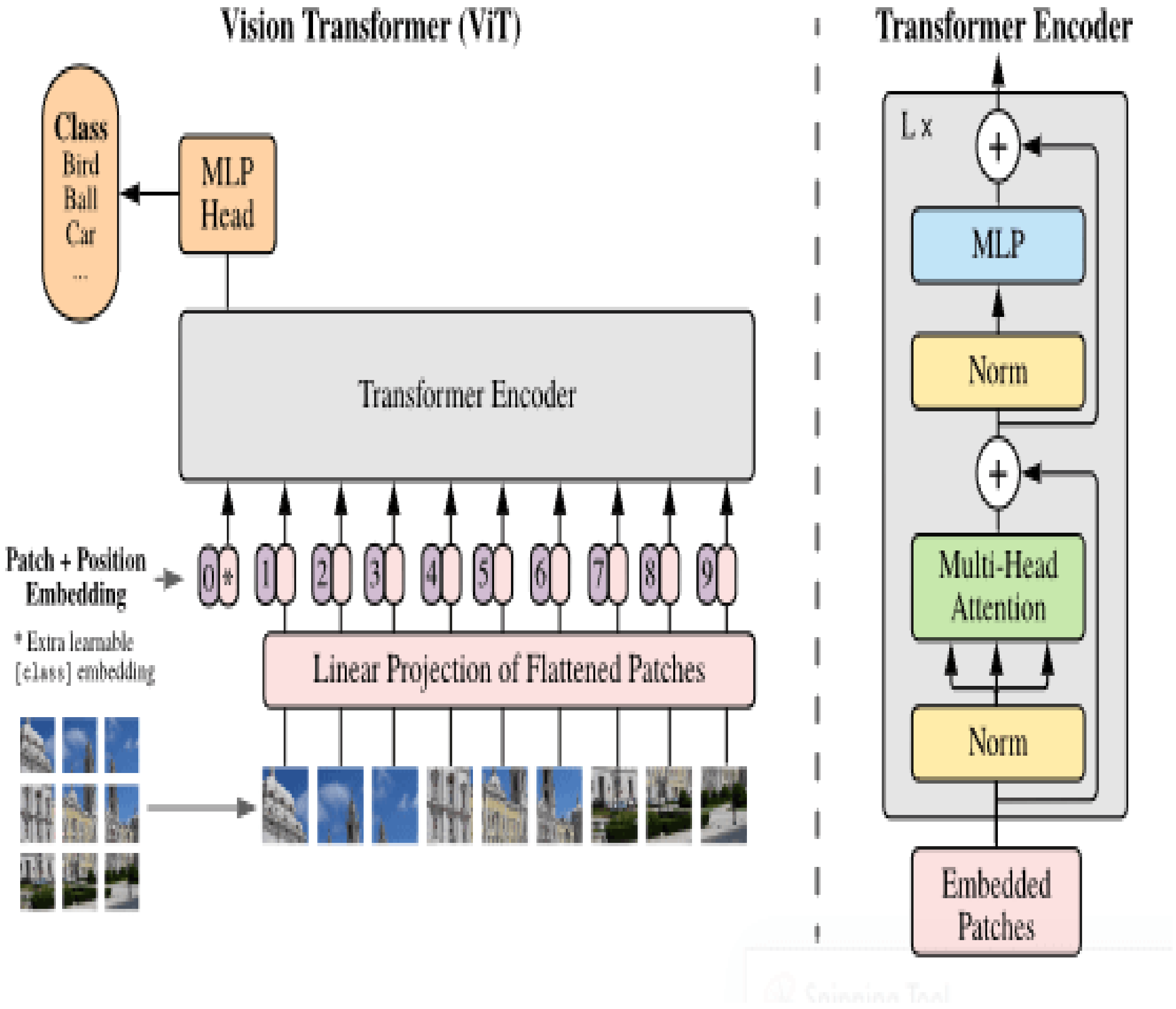


FIGURE 3.6: Vision Transformation Architecture, adopted from [viso.ai]

3.6 ENSEMBLE

Ensemble learning is a general meta approach to machine learning or deep learning that seeks better predictive performance by combining the predictions from multiple models. The ensemble model we are using in this project is Majority Voting Ensemble.So we are going to get the output predictions from our different model such as VGG19,EfficientNet,vision Transformer and then the result will be predicted by voting the prediction of all three models. If prediction from two or more model is true then it will be considered as true in final prediction. The idea of ensemble is to build a model which gives better classifier prediction[8]. The overall ensembling architecture is given in the Figure 3.7.

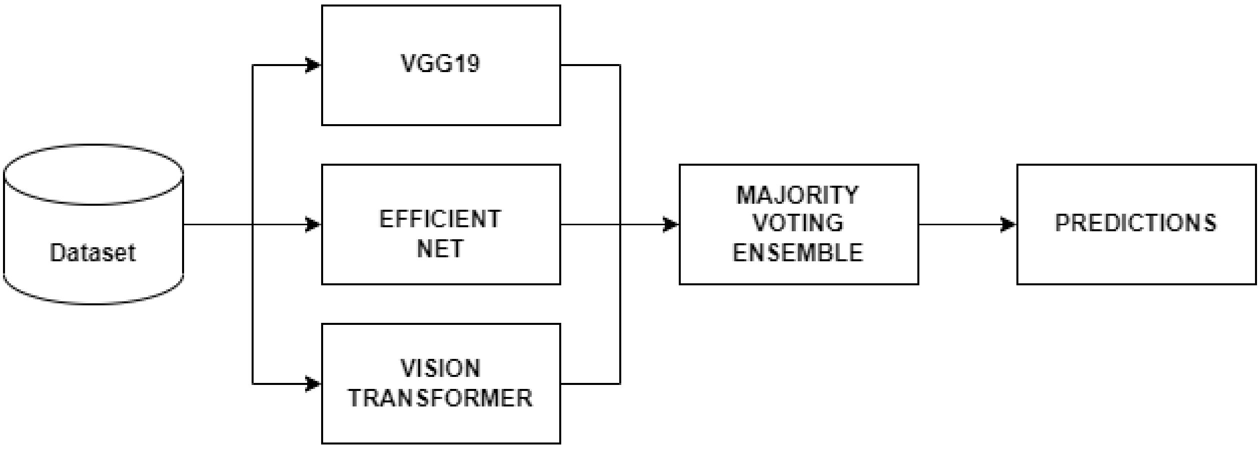


FIGURE 3.7: Ensemble Architecture

CHAPTER 4

IMPLEMENTATION

4.1 DATASET DESCRIPTION

4.1.1 Mediaeval dataset description

The Dataset was provided by Mediaeval for their online challenge on visual sentiment analysis 2021. The dataset is a collection of natural disaster images and the sentiment related to the images which can be used for visual sentiment analysis works. The given dataset contains two types of labels for two different classification.The dataset has

- Single label dataset
- Multi label dataset

4.1.1.1 Single label dataset

In a single label, each image contains one label. The labels given in the single label are Negative, Neutral and Positive. Total number of images given in the dataset is 2432. The number of images each label has are given below.

Label Name	No of Images
Negative	1695
Neutral	648
Positive	89

TABLE 4.1: single label dataset description

4.1.1.2 Multi label dataset

In multi labels, each image contains at least 3 labels. The labels mentioned in the multi label are anger, anxiety, craving, emphaticPain, fear, horror, joy, relief, sadness and surprise. Total number of images in the dataset for multi labels is 2432. The labels as well as the number of images containing that label are given below.

Label Name	No of given sentiments in all the Images
Anger	564
Anxiety	972
Craving	140
Emphatic Pain	814
Fear	1197
Horror	583
Joy	402
Relief	277
Sadness	1839
Surprise	469

TABLE 4.2: Multi label dataset description

These are two different types of labels given in the dataset which were used in our visual sentiment analysis project.

Image	Single label	Multi label
	• Positive	• Joy • Relief

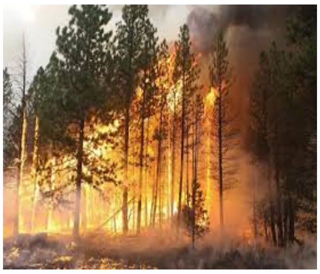
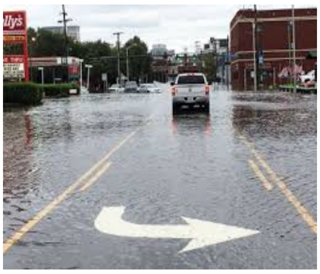
	• Negative	• Anger • Anxiety • Emphatic pain • Sadness
	• Neutral	• Anger • Anxiety • Fear

TABLE 4.3: Samples images with different labels in the dataset

4.1.2 Other dataset

In this project we have also used other dataset to check how our model works for them.

4.1.2.1 cornell dataset

This dataset was collected from kaggle. This was created for academic purpose for 'A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions'. Cornell dataset consists images related to visual sentiments. The images in cornell dataset were crawled from social media platforms. It was annotated by different people based on their sentiments and finally evaluated all the annotation and classified them accordingly into 6 classes. Cornell dataset is a multi class dataset with 6 classes such as Joy, anger, disgust, fear, sadness and surprise. The dataset consists a total of 1980 images.

Label Name	No of given sentiments in all the Images
Anger	330
Fear	330
Disgust	330
Joy	330
Sadness	330
Surprise	330

TABLE 4.4: Multi class cornell dataset description

4.2 DATA PRE-PROCESSING

4.2.1 Data Pre-Processing for mediaeval dataset

4.2.1.1 Data Pre-processing for single label

The images were read from the dataset and the read images were resized to 128x128x3(width, height, number of RGB channels). Then the images were converted into a numpy array so that it can be given as an input for our model and its corresponding labels were stored in another array. We found that the labels in the dataset for single labels are imbalanced which affects the training of the model.

So to overcome this problem we have oversampled the images using SMOTE to make a balanced dataset which adds sampled images to labels with less images that are generated from original images of the related labels. After sampling, each label contains an equal number of images which is 1695. For SMOTE k-neighbours we have used is 2 and a random state of 50. SMOTE combines the values of the neightbour images and combine them to produce a image

Early Stopping method also included to stop the model training when there is no improvement in learning. This is included to stop the model from learning differently.

The dataset is split into two halves. 80% of the images are used for training the model, 10% of the image is used for the validation of the model and the remaining 10% is used for the testing of the model.

4.2.1.2 Data Pre-processing for multi label

The images were read from the dataset and the read images were resized to 64x64x3. The images were converted into a numpy array.Then its corresponding labels were from the csv files and stored in the separate array. Now these arrays can be given as the input. Since it is multi label images, oversampling can't be done.

Early Stopping method also included to stop the model training when there is no improvement in learning. This is included to stop the model from learning differently.

The dataset is split into two halves. 80% of the images are used for training the model, 10% of the image is used for the validation of the model and the remaining 10% is used for the testing of the model.

4.2.1.3 Data pre-processing for cornell dataset

The images were read from the dataset and the read images were resized to 128x128x3. The images were converted into a numpy array.Then its corresponding labels were stored in the separate array. Now these arrays can be given as the input.

The dataset is split into two halves. 80% of the images are used for training the model, 10% of the image is used for the validation of the model and the remaining 10% is used for the testing of the model.

4.3 EXPERIMENTATIONS

4.3.1 Experiments With Single Label Dataset

4.3.1.1 VGG-19

VGG 19 pre-built model was imported.Then its input size was changed to 128x128x3 and imagenet weights were used for the model. Now the model has changed to support our project. In the model input we changed the trainable layer to false to freeze all the layers so it can use the same weights for all layers and not the updated ones. In the model output we are using flatten layers to flatten the input to give a one dimensional output which we use as an input in the next dense layer to classify the output. A dense layer is also added and in this number classes is mentioned as 3, so it can give only three output classifications which we use to classify the sentiment that the image displays. In the same dense layer, we will be using softmax as an activation function since it is single label classification. While compiling the model, we used Sparse Categorical Cross Entropy since one only label belongs to the image out of three labels. And also Adam as optimizer which is a stochastic gradient descent method. The Adam optimizer is based on the adaptive estimates of first and second order moments.

We ran the model with the unbalanced dataset with batch size as 32 and found there is a drop in learning,training accuracy as well as its testing accuracy also its outputs were recorded. In order to improve its accuracy as well as learning we used a sampled dataset which showed good learning and also good accuracy. The results obtained from the model using sampled dataset were noted. After training the model,the model is validated with a validation set and its results are also noted. The training and validation loss of this model is given in the Figure 5.2.

4.3.1.2 EfficientNet

The EfficientNet model was built from scratch. In the EfficientNet model input shape is fixed as 128x128x3. Then three blocks of layers were included in the model. Each block has a 2 convolutional, 2 depth wise convolutional and a max pooling layer. Also on top of each layer we included a Batch Normalization layer and a leaky RELU layer. In the model output, the flatten layer is included to convert the input into a 1 dimensional array and also a dense layer is added which is used to give only the number of classifications which we wanted based on the number of classes. In our experiment we will have a number of classes as 3. Softmax is used as the activation function in dense layers. In model compiling we used the same sparse categorical cross entropy as loss function and Adam as the optimizer.

Like VGG19 first we ran the model with an unbalanced dataset with batch size as 32 and obtained average results. Then we ran the same model with the sampled dataset and obtained better results. The results obtained from the model noted for further study. After training the model,the model is validated with a validation set and its results are also noted. The training and validation graph for efficientnet in Figure 5.3.

4.3.1.3 Vision Transformer

The Vision transformer model was also built from scratch. For this model, input shape is fixed as 128x128x3. We are using 16 as patch size for the model. Based on the input size and patch size, we calculated the number of patches. After calculating the number of patches, the patches were separated. Then the patches were encoded with positional values. Once the encoding is finished then the output obtained from encoding is normalized which then passed into multi head attention to vectorize the inputs. Then it's passed through the mlp which consists of a dense to convert an input into 1 dimensional output and also a dropout layer to randomly delete the data from the output 0.5% everytime. The use of mlp is used to find the hidden features of the images. The output from mlp is passed through the normalization layer to normalize the output before passing it into the next layer.Each normalization layer has an epsilon of 1e-6. A final dense layer is added which is used to reduce the output features into 3 classifications which can be used to classify the dataset. In model compiling we used the same sparse categorical cross entropy as loss function and Adam as the optimizer.

We ran the vision transformer model with an unbalanced dataset and found that the results are average like the previous two models. In this model, we have a set the learning rate as 0.001 and weight decay as 0.0001. And also we used 8 transformation layers with 32 as batch size.We have also given the projection dimension as 64 so that it can extract features from wider angles. The number of heads used for this project is 8.Then the model is run with the sampled dataset and obtained better results. The results are noted. After training the model,the model is validated with a validation set and its results are also noted.The training and validation graph for vision transformer in Figure 5.4.

4.3.1.4 Ensemble

For Ensembling we have created a majority voting ensemble. For this voting ensembler, we have loaded all the trained models of VGG19, EfficientNet and VisionTransformer. After that these three loaded models were included in an array. First we displayed the F1 score and accuracy of each model separately. Then we displayed the F1 score and accuracy of ensembled model by summing the classification values from the models and then getting the maximum value out of their values. The ensemble results were obtained only for the models trained with sampled dataset.Based on the maximum value, the label for the image is found. Finally the ensemble results were noted down. For this ensemble model input shape of image should be 128x128x3.

4.3.2 Experiments With Multi Label Dataset

4.3.2.1 VGG 19

The VGG19 pre-built model was imported from tensorflow packages. This model input shape was fixed as 64x64x3 and input weight imagenet was included. The include top is given as false since our model has to extract features from images to classify. We also changed the model input by making the trainable as false. For the output, the flatten layer is included to convert the classification output into a 1 dimensional array. This 1 dimensional array is passed through a dense layer with 3076 neurons and relu as activation function to make classification before passing it into the last layer. In the final dense layer, the number of classes is mentioned as 10 with sigmoid as activation function. Sigmoid function is used to make sure that the output will be 0 and 1. While compiling the loss function used is binary cross entropy since it is widely used for multi label classification and Adam as the optimizer.

Then the model is trained with dataset images and labels and the batch size is given as 32. The results obtained from the model are noted. Finally the model is validated with a validation set and its results are also noted for further study. The output of multi label classification like probabilities for all classes there will be no definite 1 or 0. So we used a function to convert the values with a standard threshold value of 0.5. If it is greater than 0.5, the value is changed to 1 or the value is changed to 0. This model gives the label names for array indexes whose array values are 1.Those labels are found to be the sentiment related with the images based on the output given by the model.The training and validation loss of this model is given in the Figure 5.6.

4.3.2.2 EfficientNet

For this multi label classification also, the EfficientNet was built from scratch. The input shape for this model is also fixed as 64x64x3. Then three blocks of layers were included in the model. Each block has a 2 convolutional, 2 depth wise convolutional and a max pooling layer. Also on top of each layer we included a Batch Normalization layer and a leaky RELU layer. In the model output, the flatten layer is included to convert the input into a 1 dimensional array and also a dense layer is added. This dense layer was with 4096 neurons and relu as activation function to make a hidden classification before passing it into the final dense layer. The final dense layer is used to make the final classifications which we wanted based on the number of classes. The number of classes is given as 10 and the activation function is given as sigmoid. The model is then compiled with binary cross entropy as loss function and Adam as optimizer.

Then the model is trained with the dataset and the batch size for this training is 32. The results obtained from the model are noted. Atlast the model is validated with a validation set and the results obtained are noted down. The output values are converted with a standard threshold value of 0.5. If it is greater than 0.5, the value is changed to 1 or the value is changed to 0. This model gives the label names for array indexes whose array values are 1.Those labels are found to be the sentiment related with the images based on the output given by the model. The training and validation loss of this model is given in the Figure 5.7.

4.3.2.3 Vision Transformer

The Vision transformer model was also built from scratch. For this model, input shape is fixed as 64x64x3. We are using 16 as patch size for the model. Based on the input size and patch size, we calculated the number of patches. After calculating the number of patches, the patches were separated. Then the patches were encoded with positional values. Once the encoding is finished then the output obtained from encoding is normalized which then passed into multi head attention to vectorize the inputs. Then it's passed through the mlp which consists of a dense to convert an input into 1 dimensional output and also a dropout layer to randomly delete the data from the output 0.5% everytime. The use of mlp is used to find the hidden features of the images. The mlp neuron values were fixed as [3076,2048]. The output from mlp is passed through the normalization layer to normalize the output before passing it into the next layer.Each normalization layer has an epsilon of 1e-6. A final dense layer is added which is used to reduce the output features into 10 classifications which can be used to classify the dataset. In model compiling we used the binary cross entropy as loss function and Adam as the optimizer.

We ran the vision transformer model with the dataset and found the results. In this model, we have a set the learning rate as 0.001 and weight decay as 0.0001.

And also we used 8 transformation layers with 32 as batch size.We have also given the projection dimension as 64 so that it can extract features from wider angles. The number of heads used for this project is 8. The results are noted. After training the model,the model is validated with a validation set and its results are also noted. The output values are converted with a standard threshold value of 0.5. If it is greater than 0.5, the value is changed to 1 or the value is changed to 0. This model gives the label names for array indexes whose array values are 1.Those labels are found to be the sentiment related with the images based on the output given by the model.The training and validation loss of this model is given in the Figure 5.8.

4.3.2.4 Ensemble

The majority voting ensemble is used here also. For this voting ensembler, we have loaded all the trained models of VGG19, EfficientNet and VisionTransformer. After that these three loaded models were included in an array. First we displayed the F1 score and accuracy of each model separately. Then we displayed the F1 score and accuracy of ensembled model by summing the classification values from the models and with values obtained from summing we convert the values to 0 or 1 with the threshold value of 0.7 since it is a multi label classification.

The ensemble results were obtained only for the models trained with the dataset.This ensemble model gives the label names for array indexes whose array values are 1.Those labels are found to be the sentiment related with the images based on the output given by the Ensemble model. Finally the ensemble results were noted down. For this ensemble model input shape of image should be 64x64x3.

These are the 8 experiments that are done with both the single label and multi label dataset of mediaeval. All the results obtained from these experiments are studied in the results chapter.

4.3.3 Experiments With Other Dataset

All the models that are built for Mediaeval dataset experiments were used in this experiments. we have used a standard size of 128x128x3 for all the models. Since the dataset is a multi class image classification dataset, the model was compiled with categorical cross entropy loss and softmax as activation function. Then all the model is trained for other dataset with a batch size of 32.

The results obtained from the other dataset models is noted for further study. After training the model,the model is validated with a validation set and its results are also noted.

The ensemble model is also used to improve the result of the cornell dataset. We have also used the same majority voting ensemble. Based on the result from all the models, we have used the voting method to give a finalized single output.The training and validation graph for all the models is given in the Figure 5.9.,5.10. and 5.11.

CHAPTER 5

RESULT ANALYSIS

5.1 PERFORMANCE METRICS

Our visual sentiment analysis model gives classification scores as output along with precision, recall, F1 score and accuracy. The classification scores can be used to predict the sentiment related to the images.

True Positive (TP) is the number of sentiments correctly classified Positive classes, False Positives (FP) is the number of sentiments incorrectly classified and False Negative (FN) is the number of unclassified sentiments. True Negative(TN) is the number of sentiments correctly classified in negative classes.

Precision determines the percentage of correct predictions over all the detections. The precision formula is

Precision = True Positives / (True Positives + False Positives)

Recall is the measure of all the correct positive predictions. The precision and recall will be in the range of 0 and 1.

Recall = True Positives / (True Positives + False Negatives)

F1 score is the combining of precision and recall of a classifier into a single metric by using their harmonic mean.

F1 score = (2 * Precision * Recall) / (Precision + Recall)

Accuracy is a measure of how often the classifier correctly predicts. Accuracy is defined as the ratio of number of correct predictions to total number of predictions.

Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)

Confusion matrix is used to show prediction of classification problems. The confusion matrix shows the way in which your classification model is confused while making predictions.

5.2 RESULTS FOR SINGLE LABEL DATASET

5.2.1 Without sampling dataset results

The results obtained from the three models using the original dataset are given in the Table 6.1.

TABLE 5.1: Results obtained from the models using dataset without sampling

Model	Precision	Recall	F1 Score
VGG19	0.74	0.69	0.71
EfficientNet	0.82	0.65	0.72
Vision Transformer	0.76	0.68	0.69

Though the results from the three models using the original dataset looks average but due imbalance in the dataset we found that the classifier predicts mostly the label whose input image count is higher. So we found that there are false positives in predictions. Based on the confusion matrix we also found that classifier prediction is not good. Because the labels with lesser input for learning are totally predicted wrongly. So based on the observation from the three models,VGG19 and EfficientNet classifier prediction is mostly towards negative and less towards neutral and positive labels but Vision Transformer classifier prediction is only negative labels not other labels. Because of this these three models are not good and they need to be improved.

5.2.2 Sampled dataset results

The results obtained from the three models using the sampled dataset are given in the Table 6.2.

TABLE 5.2: Results obtained from the models using sampled dataset

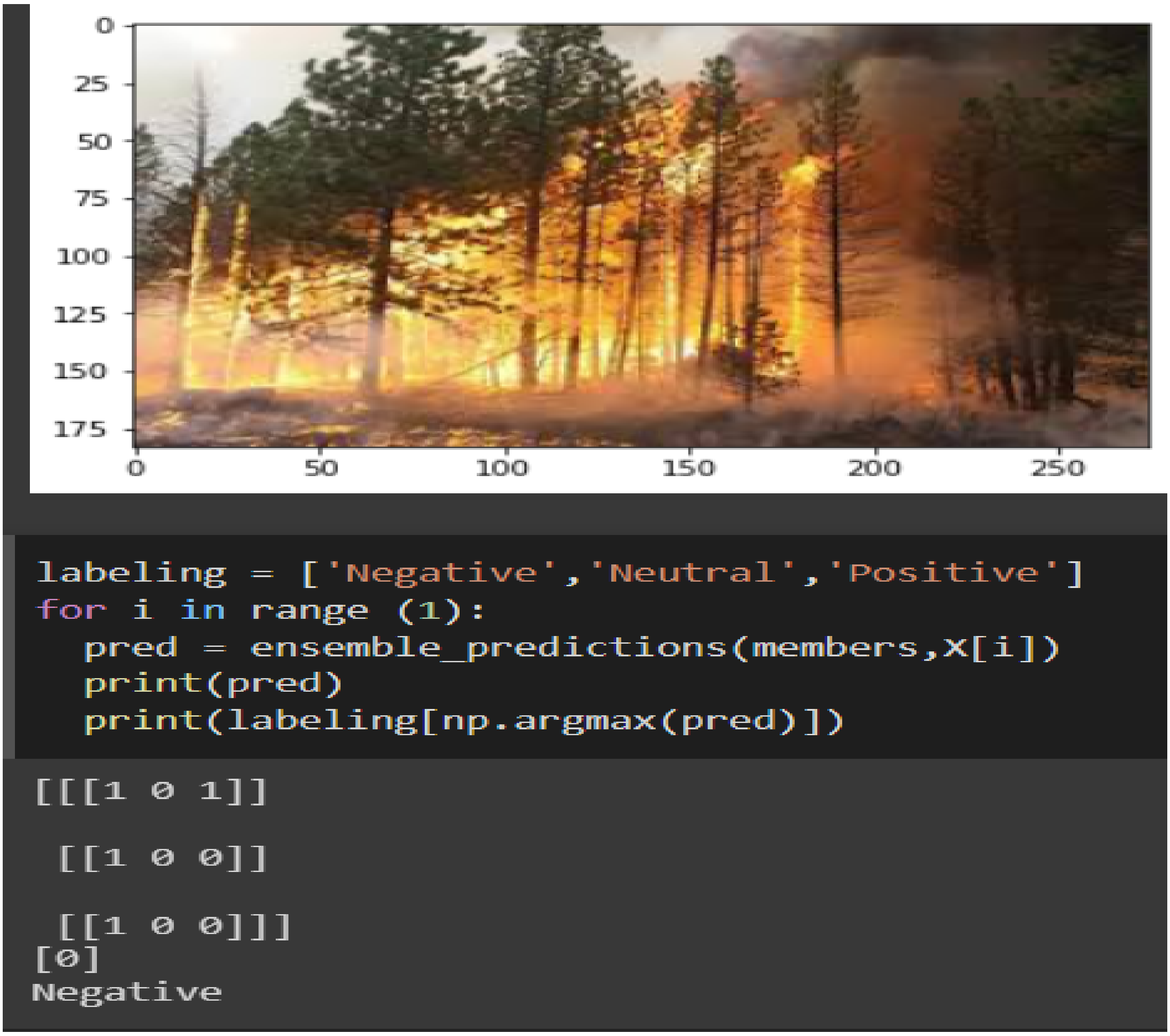
Model	Precision	Recall	F1 Score
VGG19	0.83	0.83	0.83
EfficientNet	0.87	0.87	0.87
Vision Transformer	0.82	0.80	0.81

These results from the sampled dataset are compared with the results obtained from the original dataset without sampling. We found that the precision, recall and f1 score looks way better than the results obtained from the original dataset. And also based on the study from the confusion matrix that the classifier predicts correctly for most of the images and the remaining images comes under the false positives, this results looks better for the confusion matrix also. The sampling improves the classifier prediction for other labels and it makes classifier prediction equal for all the labels. So the classifier prediction also looks good for the test set. Since these results are good. We used these models trained with sampled dataset for ensembling.

TABLE 5.3: Ensemble result obtained for single label dataset

Model	**Precision**	**Recall**	**F1 Score**
Ensemble	0.86	0.85	0.85

The Ensemble model shows a little improvement in metrics and since its results are an ensemble of three models. The classifier prediction will be more generalized and the output classification looks great. The confusion matrix of the ensemble model is also better than the confusion matrix of separate classification of these three models.



```
labeling = ['Negative','Neutral','Positive']
for i in range (1):
  pred = ensemble_predictions(members,X[i])
  print(pred)
  print(labeling[np.argmax(pred)])
```

```
[[[1 0 1]]

 [[1 0 0]]

 [[1 0 0]]]
[0]
Negative
```

FIGURE 5.1: Single label Ensemble Prediction obtained for the given image

The above image shows the result of our ensemble model for single label images. In the results where 1 are the labels present in the images. In the last output column first output line shows the result from vision transformer, second output line shows the result from Efficient Net and the third output line shows the result from VGG19. The final line shows the overall final voted output index value which then used to show the label name in the last line.

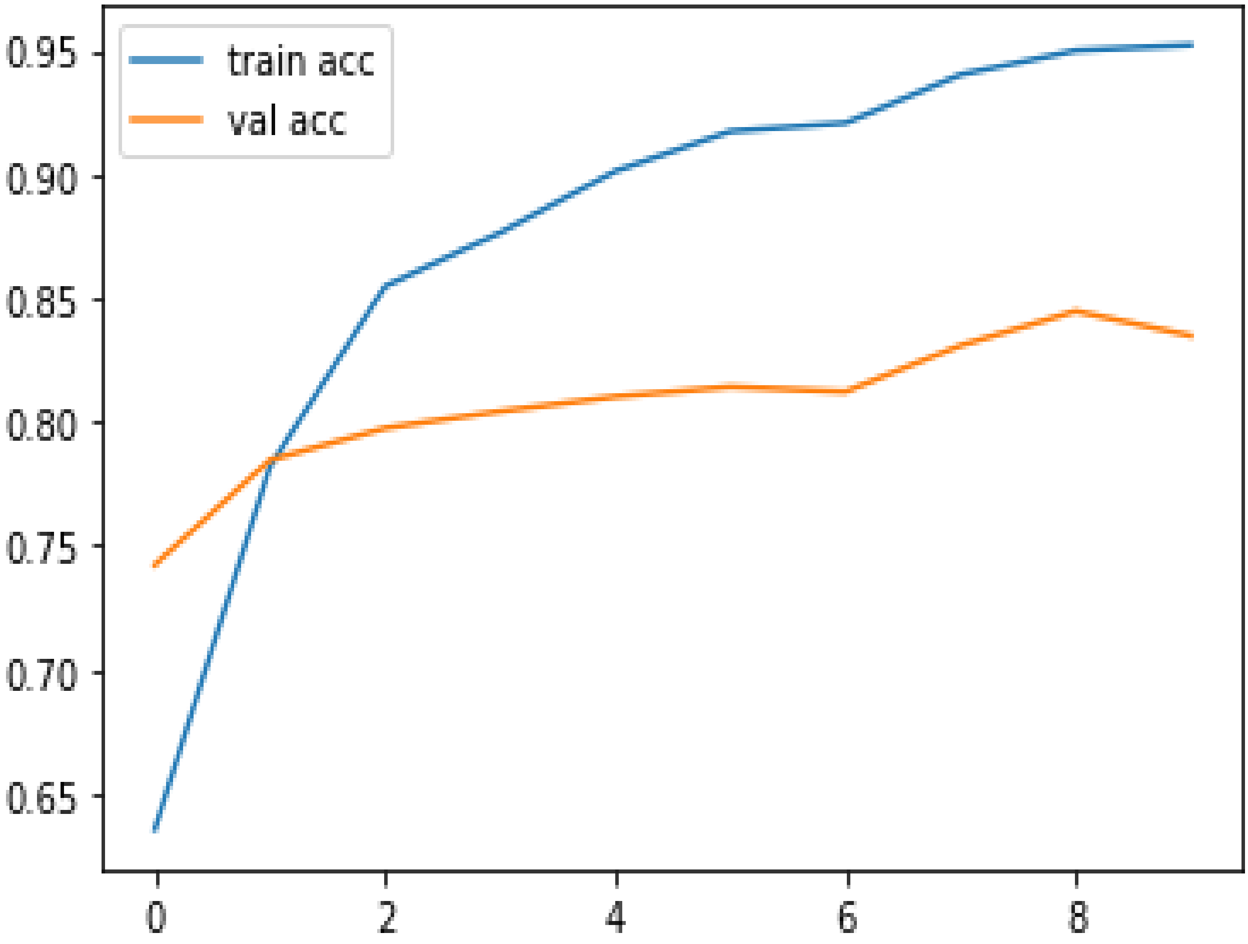


FIGURE 5.2: VGG19 training and validation accuracy graph for single label

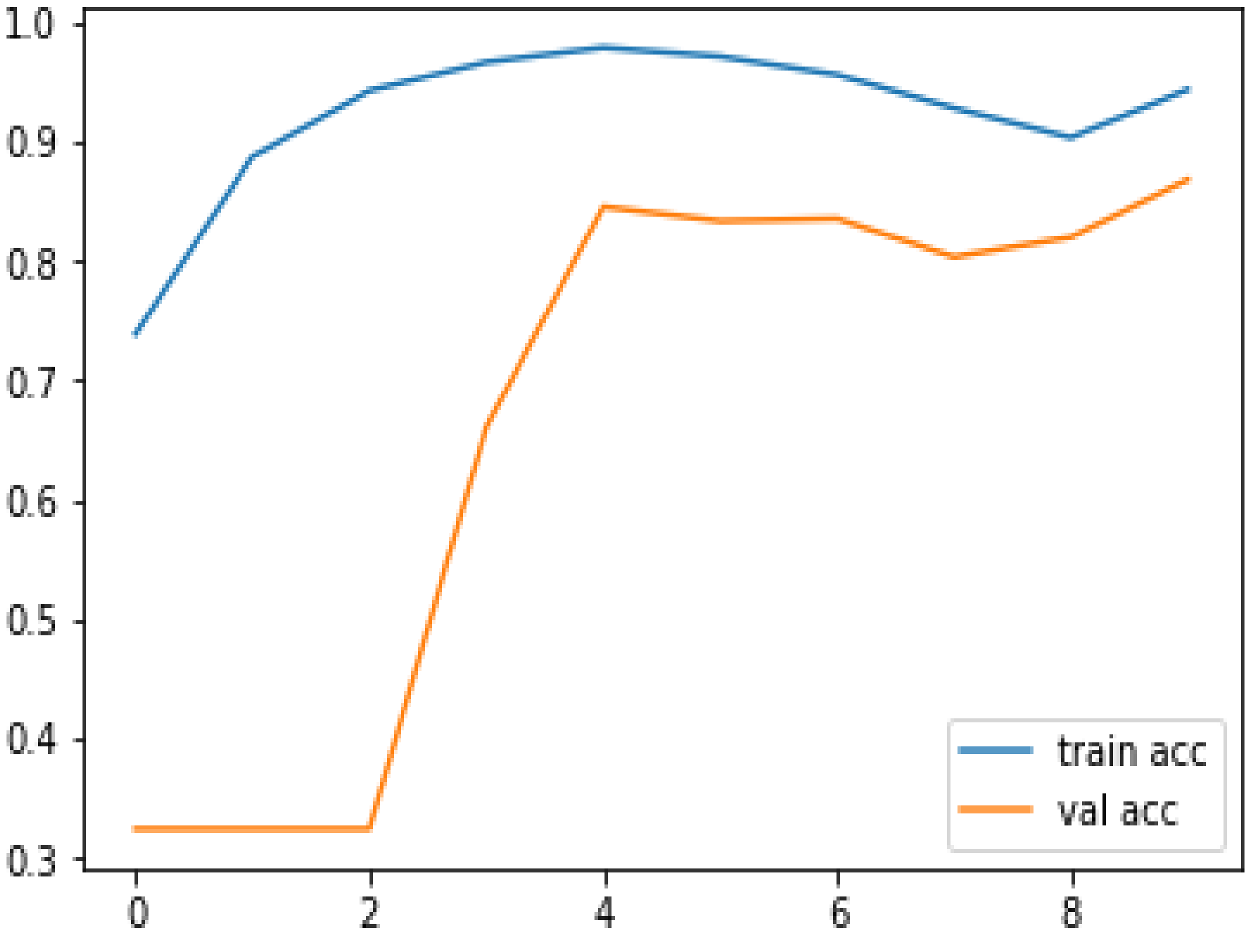


FIGURE 5.3: Efficient training and validation accuracy graph for single label

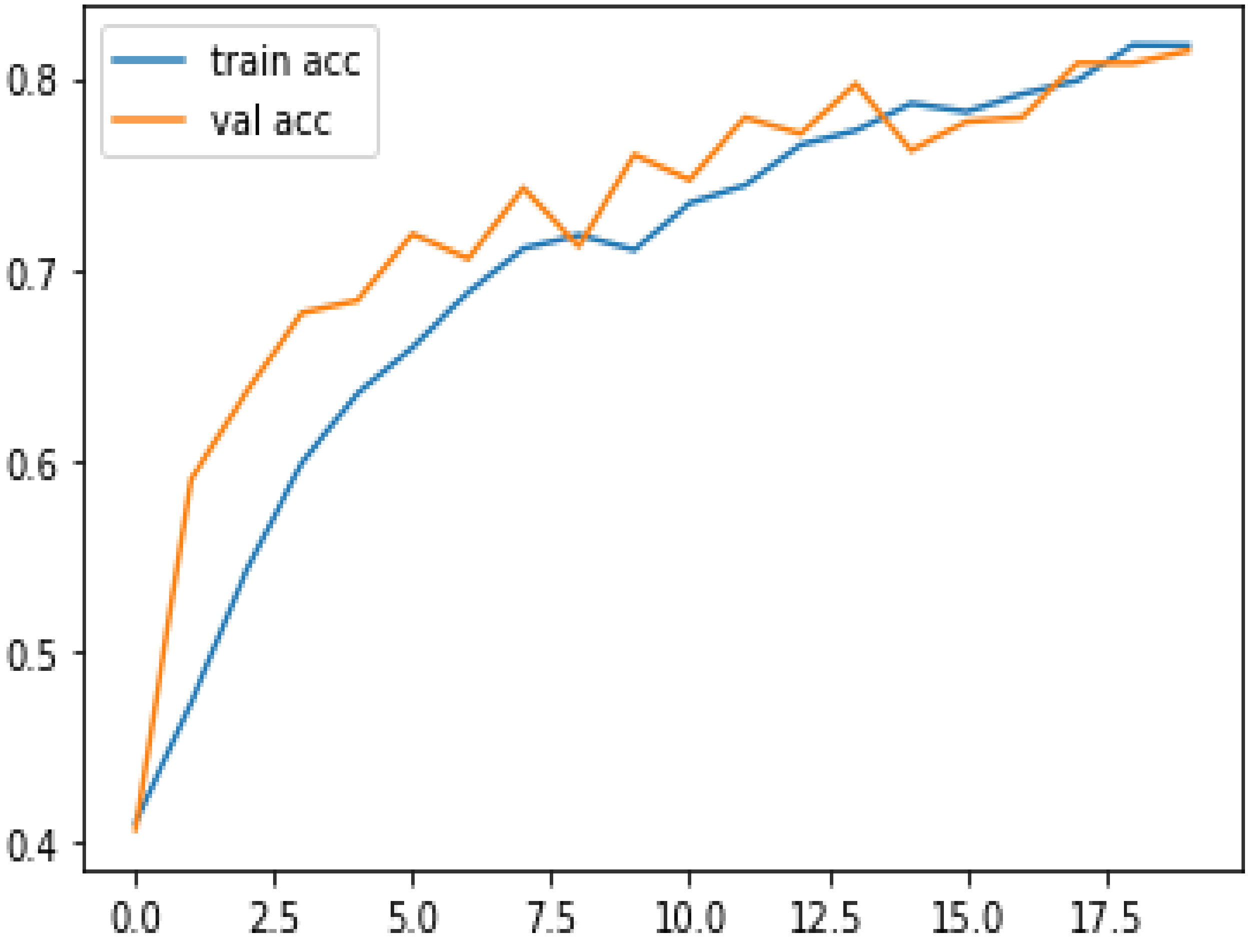


FIGURE 5.4: Vision Transformer training and validation accuracy graph for single label

The above graph shows the training and validation accuracy of their respective models. x axis shows no of epochs where 0 is considered epoch 1 and y axis shows the accuracy of the training and validation set.

5.3 RESULTS FOR MULTI LABEL DATASET

The results obtained from the three models using multi label dataset are given in the Table 6.4.

TABLE 5.4: Results obtained from the models for multi label dataset

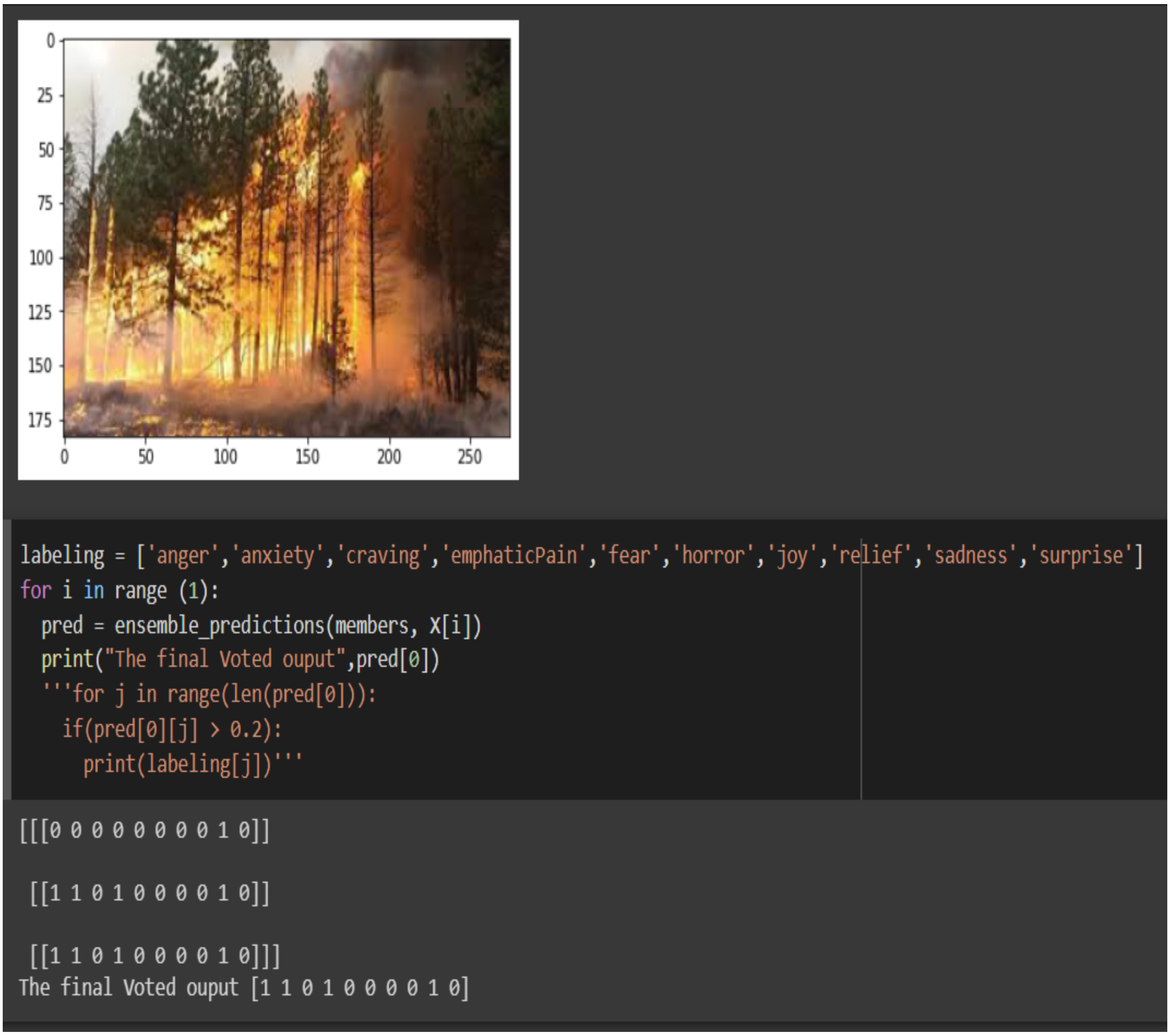
Model	Precision	Recall	F1 Score
VGG19	0.54	0.50	0.47
EfficientNet	0.57	0.42	0.43
Vision Transformer	0.52	0.67	0.55

The results show average values for Precision, Recall and F1 score. And also it is multi label classification the results metrics gonna be different from the single label classification. The confusion matrix cannot be drawn for multi label classification. Based on the study of the classification report of each model, the classification for each label looks good in both VGG19 and EfficientNet But not in Vision Transformer. In the Vision transformer model many labels were not predicted by the classifier. And also in all the three models the cravings label prediction too low since the learning for craving is also low. Since our motive is to ensemble all these models show better results. The results obtained for the same model with different threshold values and found the standard threshold value 0.5 gives better results. The ensemble result for multi label is given in Table 6.5.

TABLE 5.5: Ensemble result obtained for multi label dataset

Model	Precision	Recall	F1 Score
Ensemble	0.70	0.79	0.71

Though each model's results look average separately, our ensembled model improves the metrics and gives a better performance. The threshold value for the ensembler is different from the threshold value used for the models individually. We found a suitable threshold value as 0.7 after analyzing the results from the different threshold values. And our classifier prediction for ensemble models is also more accurate than the three model's separate predictions.



```
labeling = ['anger','anxiety','craving','emphaticPain','fear','horror','joy','relief','sadness','surprise']
for i in range (1):
  pred = ensemble_predictions(members, X[i])
  print("The final Voted ouput",pred[0])
  '''for j in range(len(pred[0])):
    if(pred[0][j] > 0.2):
      print(labeling[j])'''
```

```
[[[0 0 0 0 0 0 0 0 1 0]]

 [[1 1 0 1 0 0 0 0 1 0]]

 [[1 1 0 1 0 0 0 0 1 0]]]
The final Voted ouput [1 1 0 1 0 0 0 0 1 0]
```

FIGURE 5.5: Multi label Ensemble Prediction obtained for the given image

The above image shows the result of our ensemble model for multi label images. In the results where 1 are the labels present in the images. The array index of those 1 corresponds to the label given the labeling array index. In the last output column first output line shows the result from vision transformer, second output line shows the result from Efficient Net and the third output line shows the result from VGG19. The final line shows the overall final voted output.

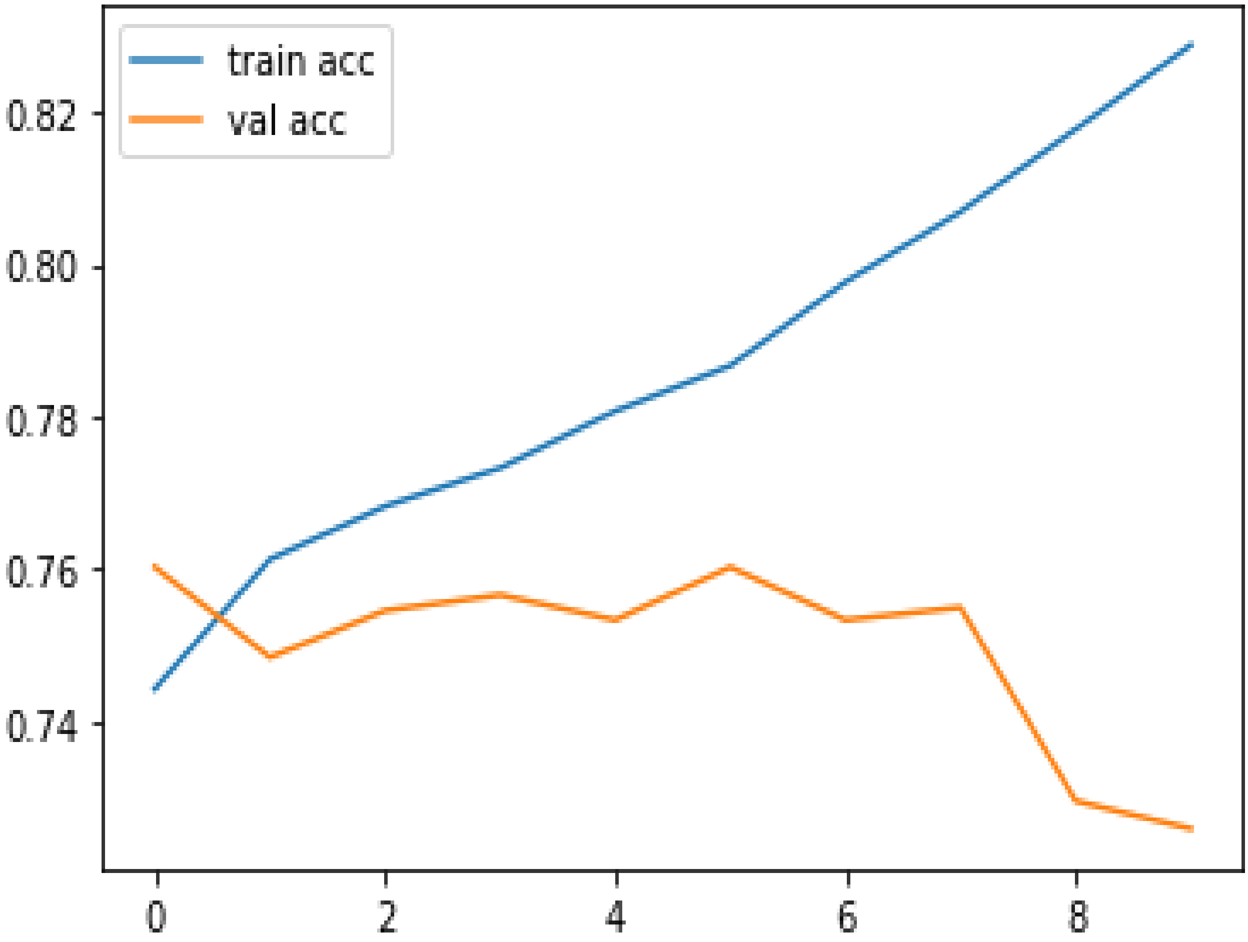


FIGURE 5.6: VGG19 training and validation accuracy graph for multi label

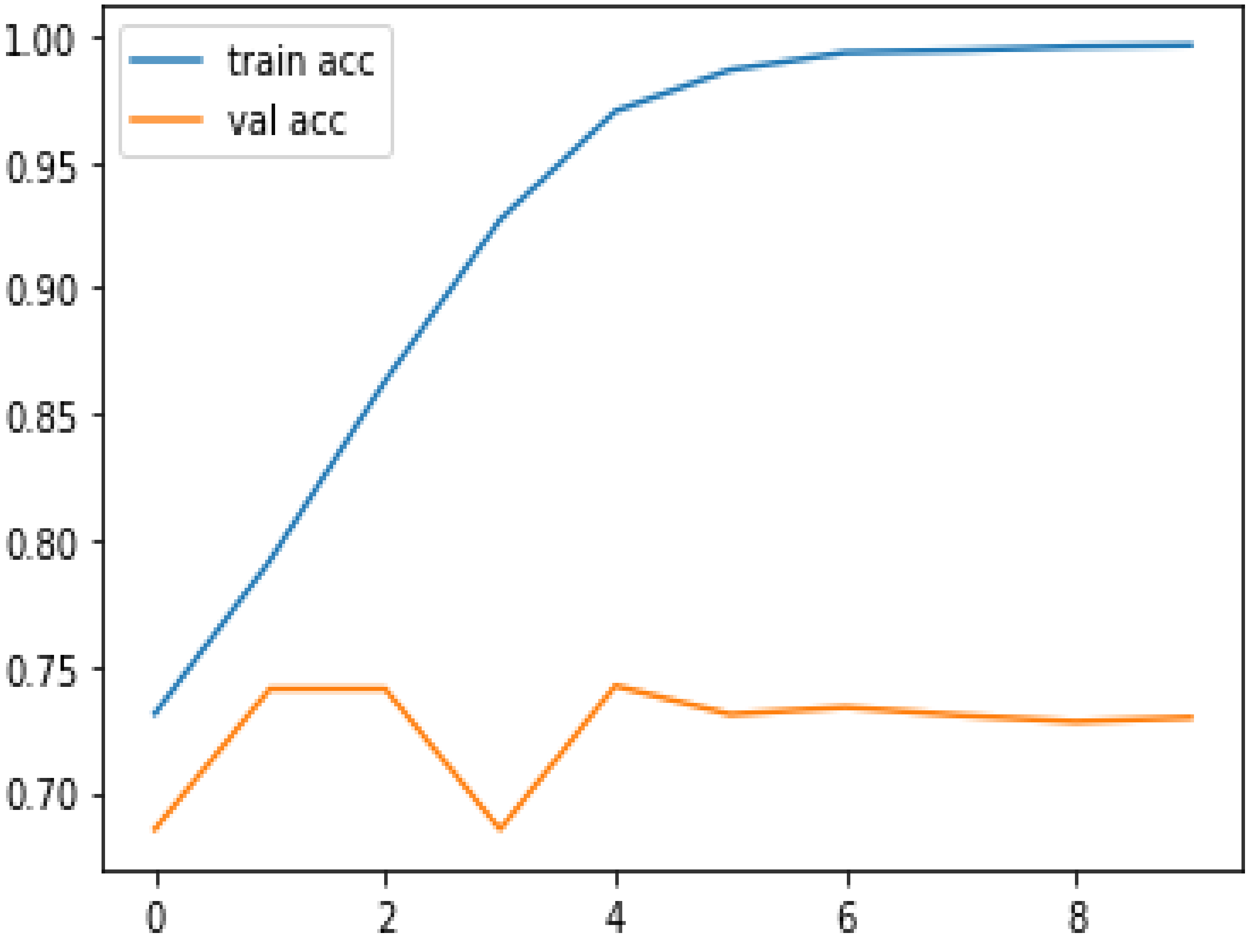


FIGURE 5.7: EfficientNet training and validation accuracy graph for multi label

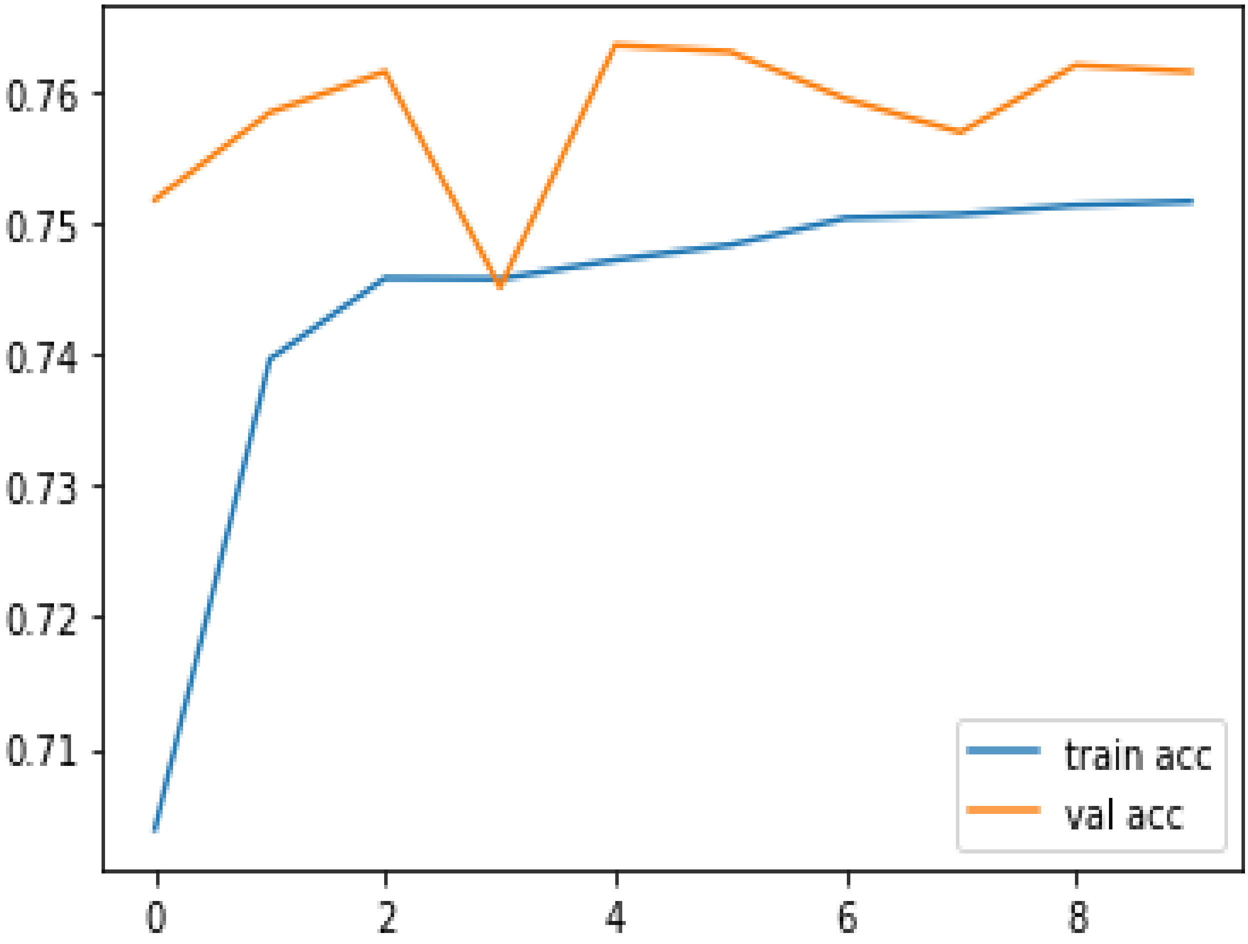


FIGURE 5.8: Vision Transformer training and validation accuracy graph for multi label

The above graph shows the training and validation accuracy of their respective models. x axis shows no of epochs where 0 is considered epoch 1 and y axis shows the accuracy of the training and validation set.

5.4 RESULTS FOR CORNELL DATASET

The results obtained from the three models using cornell dataset are given in the Table 6.5.

TABLE 5.6: Results obtained from the models for cornell dataset

Model	Precision	Recall	F1 Score
VGG19	0.46	0.42	0.43
EfficientNet	0.40	0.13	0.32
Vision Transformer	0.47	0.28	0.30

Since we used the same models that we used for single label and multi label classification for multi class classification, the results looks not good. The reason why we used the other dataset is to check how the same model for different dataset we haven't fine tuned it. And also the dataset contains 330 images for each classes when splitted 80% for training only 264 images available for training which is not enough for training because of this the result also looks not good.

TABLE 5.7: Ensemble Results obtained from the models for cornell dataset

Model	Precision	Recall	F1 Score
Ensemble	0.36	0.24	0.26

The results from separate model not looks good so the ensemble results obtained from those results also not good.

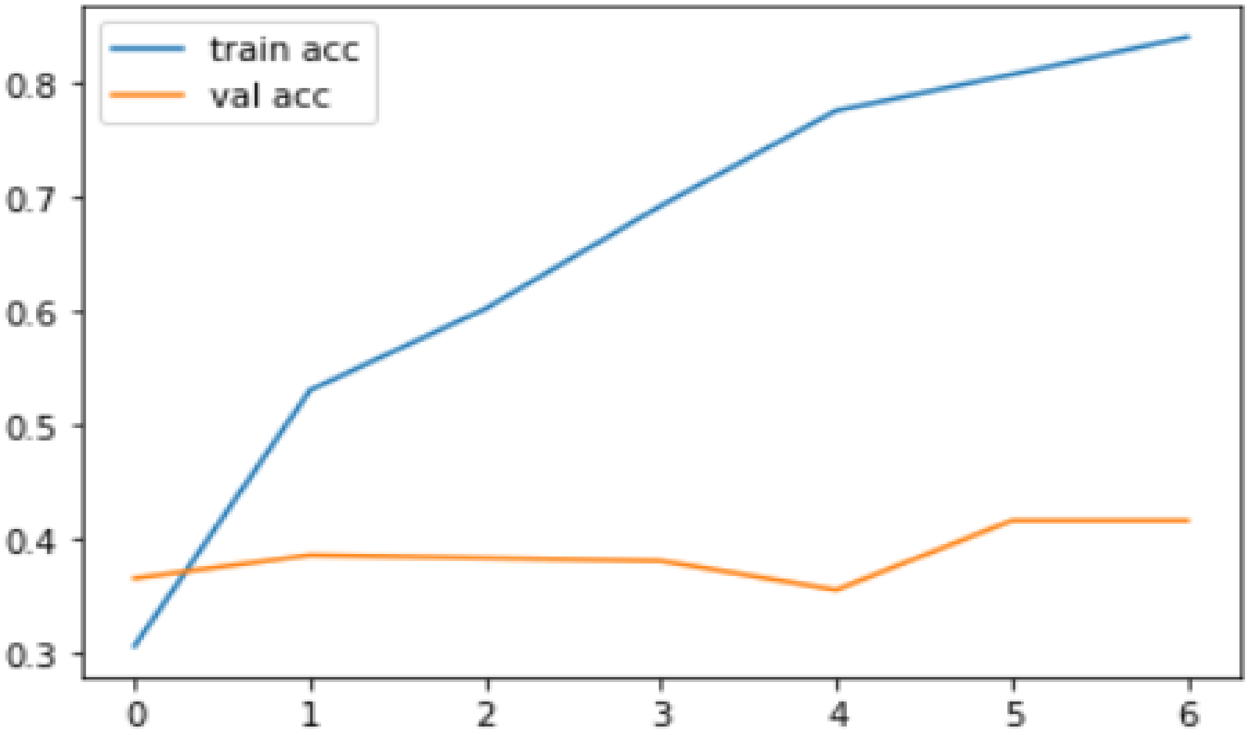


FIGURE 5.9: VGG19 training and validation accuracy graph for cornell dataset

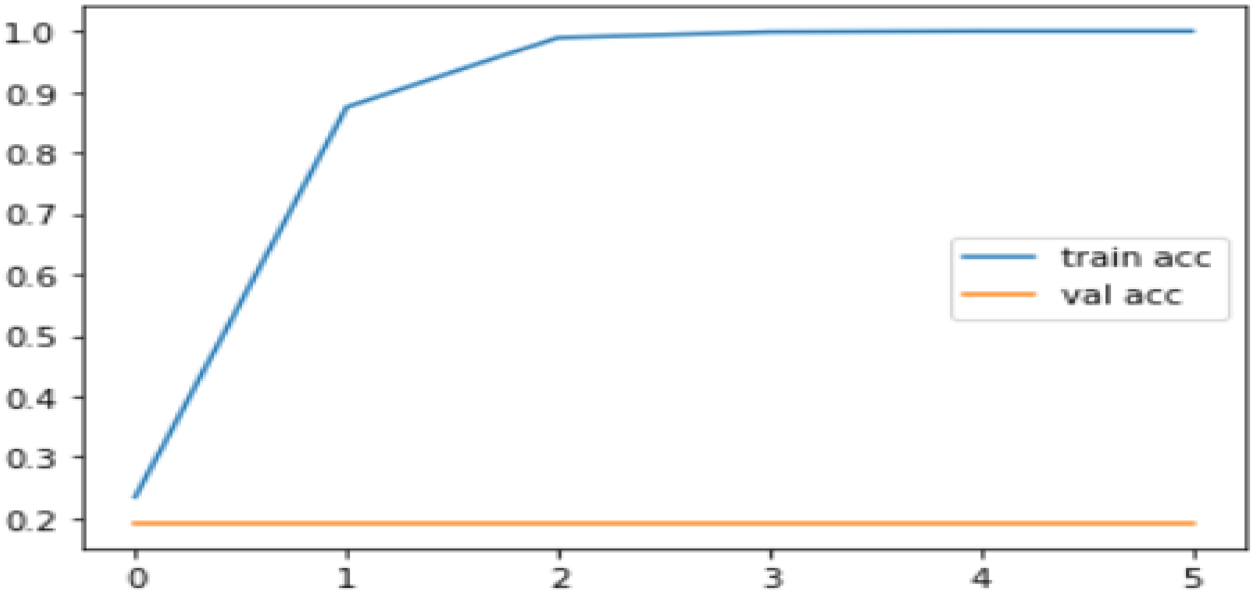


FIGURE 5.10: EfficientNet training and validation accuracy graph for cornell dataset

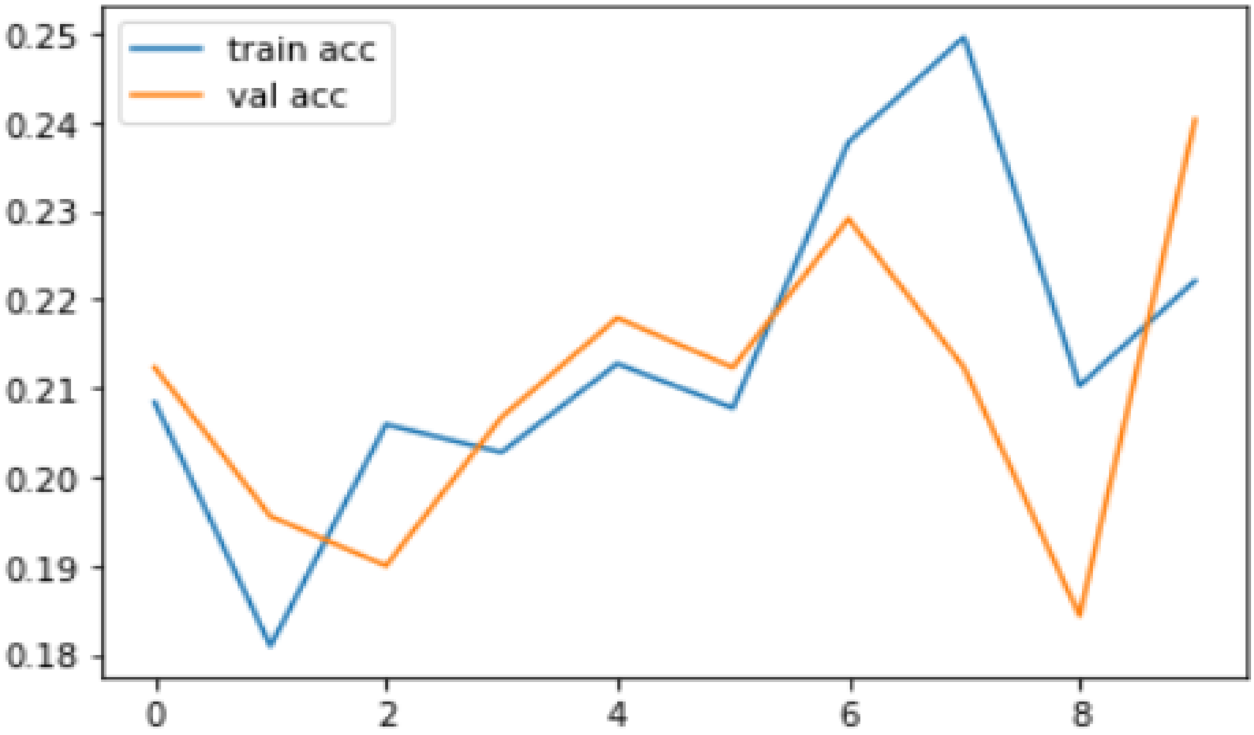


FIGURE 5.11: Vision Transformer training and validation accuracy graph for cornell dataset

The above graph shows the training and validation accuracy of their respective models. x axis shows no of epochs where 0 is considered epoch 1 and y axis shows the accuracy of the training and validation set.

CHAPTER 6

COMPARISON WITH EXISTING WORKS

We compared our results with the results submitted for the Mediaeval online challenge of visual sentiment analysis 2021 by HCMUS Teams. The HCMUS team result for the challenge is obtained from their paper “HCMUS at MediaEval 2021: Efficient methods of Metadata Embedding and Augmentation for Visual Sentiment Analysis “.

The HCMUS team used different versions of the same EfficientNet model to analyse the images of the same dataset in matlab. Whereas we used three different models which are selected based on their unique features and ensemble all the three model to create a ensemble model to give a improved results. The weighted F1 score result of their models and our model is given below.

TABLE 6.1: Result Comparison of weighted F1 Score

Model	Our Model	HCMUS Team Model
Single Label Model	0.85	0.77
Multi Label Model	0.71	0.58

As we can see, the weighted f1 score of our model looks better than the weighted f1 score of the HCMUS Team model. With these results we conclude that our model performs better than the HCMUS Team model in terms of label-wise predictions of the classifier. This shows how our models differs from their in terms of the results by using three different models for ensemble.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The test is made out of two errands including a solitary name and multi-mark picture grouping undertakings with various arrangements of marks. The primary errand means to cover the traditional three classifications/marks commonly used to address feelings. The other two errands mean to cover sets of names more intended for catastrophic events. These three arrangements of names permit to investigate various parts of the space, and the assignment's intricacy increments by going further in the feelings order. For every one of the assignments, we depend on three best in class profound structures specifically VGG-19,Efficient Net and Vision transformer. To this point, the models pre-prepared on the ImageNet dataset are calibrated on the advancement dataset.

In the ongoing executions, we depend on object-level data exclusively by utilizing the models pre-prepared on ImageNet dataset. We accept, scene-level elements could likewise add to the errand. Later on, we plan to mutually use both article and scene-level data for better execution on all the tasks. Moreover, we mean to utilize merit-based combination plans by thinking about the commitment of the singular models to the assignments.

REFERENCES

1. Alexey Dosovitskiy.; Lucas Beyer.; Alexander Kolesnikov.; Dirk Weissenborn.;Neil Houlsby.; AN IMAGE IS WORTH 16X16 WORDS:TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE.In Proceedings of the International Conference on Learning Representations(ICLR) 2021.03 June 2021.

2. Deep Models for Visual Sentiment Analysis of Disaster-related Multimedia Content. arXiv:2112.12060,30 Nov 2021

3. Ganesh Chandrasekaran, Naaji Antoanela, Gabor Andrei, Ciobanu Monica, Jude hemanth. Visual Sentiment Analysis Using Deep Learning Models with Social Media Data.Special Edition- User Experience for advanced Human Computer Interaction. 2022

4. Hai Ha Dohaiha, PWC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: a comparative review. Expert Systems with Applications, 2018.

5. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

6. Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1253, 2018.

7. Sicheng Zhao, Hongxun Yao, You Yang, and Yanhao Zhang. Affective image retrieval via multi-graph learning. In Proceedings of the 22nd ACM international conference on Multimedia, pages 1025–1028. ACM, 2014.

8. Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586, 2014.

9. Weining Wang and Qianhua He. A survey on emotional semantic image retrieval. In 15th IEEE International Conference on Image Processing, pages 117–120, 2008.

10. Wenjing Gao, Wenjun Zhang, Haiyan Gao and Yonghua Zhu. Visual sentiment analysis via deep multiple clustered instance learning. Journal of Intelligent Fuzzy Systems (2020) pages 7217–7231

11. Víctor Campos, Amaia Salvador, Xavier Giró-i Nieto, and Brendan Jou. Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In Proceedings of the 1st International Workshop on Affect Sentiment in Multimedia, ASM '15, pages 57–62, New York, NY, USA, 2015. ACM.