



Leveraging Transfer Learning to Enhance Real-Time Emotion Recognition

Abhinav Pavithran

Student ID: 201632512

Project Supervisor: Flávia Alves

Secondary Supervisor: Igor Potapov

A DISSERTATION

Submitted to

The University of Liverpool

in partial fulfilment of the requirements
for the degree of

MASTER OF SCIENCE

September 22, 2023

Abstract

Emotion detection is a growing field of research in computer science with a wide array of applications. It can be used in advertisement to gauge how a customers responds to specific products or in the education sector to assess student engagement with lectures and course work. In the past traditional machine learning methods were used for emotion recognition, however in recent years convolution neural networks have made significant advances in this field and prove to perform better than machine learning approaches.

However even with CNN approaches there are certain drawbacks such as computational expense, training time and dataset limitations. This project aims to resolve those drawbacks using a transfer learning approach; a pre-trained CNN model is fine-tuned by training a limited number of layers on a smaller dataset to achieve emotion recognition. As only a limited number of layers are being trained the amount of computation and training time reduces. Since the model is pre-trained it has already learned many features so it can be fine-tuned on a smaller dataset.

The goals of this project are choosing the right dataset so that the model can generalize well and choosing the right pre-trained models that can extract features well but aren't too computationally expensive. The software developed should recognize emotions in real time using computer vision for face detection and the CNN to predict the emotion.

In this project the right models were fine-tuned and tested and performed to a reasonable accuracy in real time. This project shows that by using a transfer learning approach a pre-trained CNN can be fine-tuned to predict emotions to a reasonable accuracy in relatively less time using lower computational resources.


Student Declaration

I confirm that I have read and understood the University's Academic Integrity Policy.

I confirm that I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.

I confirm that I have not copied material from another source nor committed plagiarism nor fabricated data when completing the attached piece of work. I confirm that I have not previously presented the work or part thereof for assessment for another University of Liverpool module. I confirm that I have not copied material from another source, nor colluded with any other student in the preparation and production of this work.

I confirm that I have not incorporated into this assignment material that has been submitted by me or any other person in support of a successful application for a degree of this or any other university or degree-awarding body.

SIGNATURE  _____

DATE September 22, 2023

Acknowledgments

Thank you to my supervisor, Ms. Flávia Alves, for providing guidance and feedback throughout this project. Thanks also to my flatmates Saif, Kunal, Harshitha, Jeetanshu, Saakshi and my friends Nakul, Excy and Yerkezhan for putting up with me being sat in the kitchen for hours on end, and for providing guidance and support when required.

Contents

1	Introduction	1
2	Aims and Objectives	2
2.1	Aims	2
2.2	Objectives	2
3	Background	3
3.1	Dataset	3
3.2	Transfer learning	4
3.2.1	VGG	4
3.2.2	InceptionV3	5
3.2.3	Xception	5
3.3	Real time detection	6
4	Ethical Use of Data	7
5	Design	8
5.1	Model training	8
5.2	Real time detection	8
5.3	Output	9
5.4	Original design	9
5.5	Changes to original design	9
6	Implementation	10
6.1	Data preprocessing	10
6.2	Model fine-tuning	11
6.3	Real time detection	12
7	Evaluation	13
8	Learning Points	14
9	Professional Issues	16
9.1	Public Interest	16
9.2	Professional Competence and Integrity	16
9.3	Duty to Relevant Authorities	17
9.4	Duty to the Profession	17
10	Conclusion	18
	References	18
A	Training Metrics	22

List of Figures

3.1	AffecNet dataset sample	3
3.2	FER+ dataset samples	4
3.3	VGG Structure	4
3.4	InceptionV3 Structure	5
3.5	Xception Structure	5
3.6	Caption	6
5.1	User Interface Mockup	9
6.1	Data distribution before augmentation	10
6.2	Data distribution after augmentation	10
6.3	Real time emotion detection output	12
A.1	VGG16 Metrics Graph	22
A.2	VGG19 Metrics Graph	22
A.3	InceptionV3 Metrics Graph	22
A.4	Xception Metrics Graph	23

List of Tables

3.1	Model details	5
6.1	Metrics table	11

Introduction

One aspect of being human that is common to all is having emotions. Whether you are a child, teenager, adult, or an elderly person we all possess emotions and feelings. We can't hide our emotions however hard we try sometimes. Emotions can be described as a bodily sensation and/or behavior that reflects the personal significance of an event, thing, or situation. If a computer can understand non-verbal communications from the user it opens many avenues of applications. For example, it can be used in education where a computer can assess the interest of students in class. In the medical field it can be used in psychiatry to better understand what the patient is feeling. It could also help with patients of autism as well, which can help doctors assess what state the patient is in and what treatment needs to be administered. This technology can also be used in crime interrogation to detect lies and even in ATMs to detect if the withdrawer is scared and aid if needed [19].

In the past many approaches have been implemented to achieve this. The two popular approaches are machine learning and Convolutional Neural Networks (CNNs). In this project we will be focusing on CNN approaches. CNNs are very powerful tools in image analysis as they are very good at learning and extracting features from images and converting them to usable data [2]. Over the years many types of CNNs have been tested. One of the drawbacks to this approach is training the network from scratch, which is computationally expensive and time consuming. Also, to train these models to reach reasonable accuracy a very large and diverse dataset is needed.

To solve this problem a transfer learning approach is used. A network that has been previously trained on millions of images is used as the base network which is then trained on another dataset. Since the network is pre-trained it has already learned many useful features and training it on new data will help fine tune the model [1]. Since it has already learned from previous data it does not require a large dataset [16]. The upper layers and/or lower layers are refined and tuned based on our dataset. There are many pre-trained networks that exist today such as ResNet, GoogleNet and DeepNet [19] [13]. Transfer learning leverages knowledge from previously trained models and uses it for training newer models. Using this approach real time detection of emotion is implemented.

Aims and Objectives

The primary aim is to accurately identify human emotions. Even though pre-trained networks are an elegant solution for training time and limited dataset, choosing the right network and technique is crucial for saving time and computational resources. The aims can be divided into two main categories.

2.1 Aims

1. Choosing the right pre-trained model: There are many pre-trained models out there that have varying characteristics. It is important to choose one that has good feature extracting capabilities but at the same time not be too computationally expensive to train. In the proposal only two models were proposed but 4 different models were fine-tuned and tested. The models chosen are VGG16, VGG19, InceptionV3 and Xception.
2. Model should be generalized: The model should be able to detect emotions of all sorts of people in varying lighting and environment conditions. Regardless of color, gender, race, facial structure, strong/weak lighting, head pose, the model should do a good job at detecting emotion [15]. The dataset considered for the project was not granted access to (AffectNet) so the back up dataset (FER+) had to be used. Due to this the model did not generalize as intended.

2.2 Objectives

1. Research different types of emotional models and investigate the type of CNNs they require for training. The models researched were Ekman's and valence-arousal emotional model.
2. Research the different types of pre-trained models available and see which models perform well and have lower trainable parameters. The model researched and trained were VGG16, VGG19, InceptionV3 and Xception.
3. Examine the available datasets for facial expression recognition to see which ones have a diverse set of images in varying conditions. AffectNet had the most diverse and sizeable dataset. FER+ was also considered and ultimately used for this project.
4. Develop different types CNN models to train on the data using different labels and see which performs optimally. All four models were trained and tested.

Background

3.1 Dataset

For the model to be able to generalize well it is important to train it on a diverse dataset. Initially the AffectNet dataset was considered. It is the largest dataset of affect in the wild. It was created by using emotional keywords and querying images from three different search engines. A total of 1250 search queries are used to explore the various search engines. After querying images, it is then processed and annotated. The total annotated images come up to 450,000 [15].

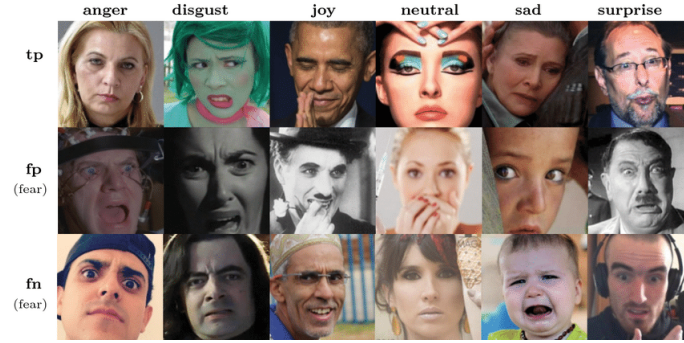


Figure 3.1: AffectNet dataset sample

This dataset measures emotions using the categorical model and valence-arousal scale. In the categorical model, emotion is chosen from a list of six basic emotions. In the valence-arousal scale it measures emotions across two axes [9]. The dimensional model can detect different kinds of emotions and measure intensity which was the drawback of other models. Valence refers to whether the emotion is positive or negative and arousal indicates the intensity whether it exciting/agitating or calm/soothing.

Another dataset that was considered is FER+ dataset. It contains 28,709 images with a resolution of about 48x48 and each image has one of 7 emotion labels. FER+ is an improved version of the FER dataset [11]. The labels were retagged using a crowd sourcing method with 10 taggers [4]. Crowd sourcing helped improve the accuracy of the tags and increasing the number of taggers to 10 helped improve the agreement between taggers to over 80%.

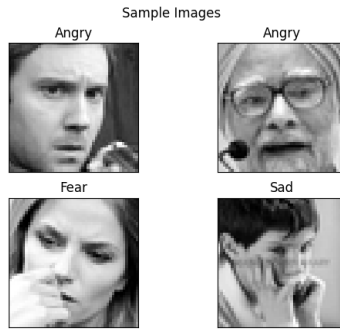


Figure 3.2: FER+ dataset samples

3.2 Transfer learning

In this project the concept of transfer learning is leveraged to obtain better results faster. Models trained on ImageNet dataset are used for fine tuning. ImageNet is one of the largest object classification dataset currently available [7] [12], models trained on this data have learned a large number features and are generalized models which make it ideal for this project. Since the network has already learned a large number of features from ImageNet it is already quite good at detecting and extracting features but the network requires fine tuning of weights in the initial and/or final layers to make it better suited for the task of emotion detection.

This setting where the target and source are different for the network is called Inductive Transfer Learning [18]. The next key aspect is choosing the right network. The network must have a decent trade-off between speed and accuracy. For this reason, InceptionV3, Xception and VGG were chosen [3], they have relatively lesser parameters to train while simultaneously having strong feature extraction and processing power [16].

3.2.1 VGG

VGG has a few different models such as VGG16 and VGG19 with the main difference in total number of layers, 16 and 19 respectively. VGG16 has 138 million and VGG19 has 143 million parameters. The network uses fixed size convolution and maxpool kernels [3]. This approach reduces the number trainable parameters which makes training faster and less computationally expensive [20].

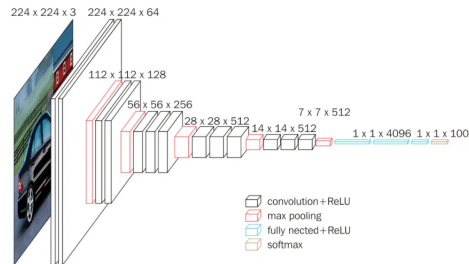


Figure 3.3: VGG Structure

3.2.2 InceptionV3

InceptionV3 is the improved version of V2 which is built by Google. It utilizes new techniques to optimize the network and improve efficiency. It has 23 million parameters and a depth of 189 layers. Utilizes factorization into smaller convolutions, spatial factorization into asymmetric convolutions, auxiliary classifiers and efficient grid size reduction techniques [21] that makes this network more efficient and powerful. Factorized and asymmetric convolutions reduce number of parameters and computational expenses. Auxiliary classifiers help the deep neural network improve convergence [21].

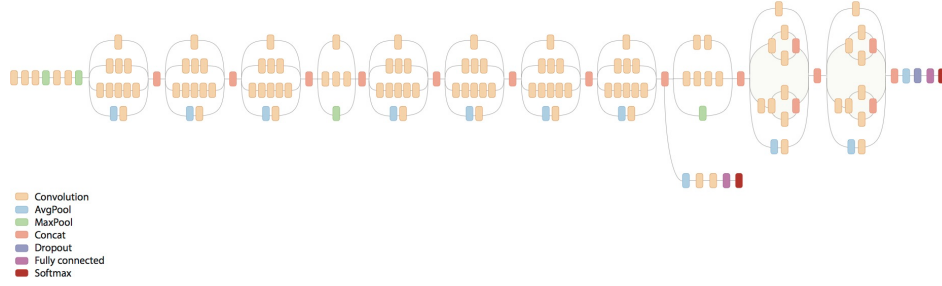


Figure 3.4: InceptionV3 Structure

3.2.3 Xception

Xception by Google is the extreme version of Inception [23]. It uses a modified depth wise separable convolution approach, this does not have intermediate RELU non-linearity [5]. The approach yields better accuracy than the Inception iterations on ImageNet database. This model has 23 million parameters and a depth of 81 layers.

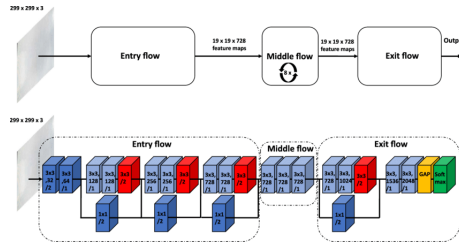


Figure 3.5: Xception Structure

The main characteristics of the above described models are summarized in the table below

Model	No. layers	Parameter size
VGG16	16	138 mil
VGG19	19	143 mil
InceptionV3	189	23 mil
Xception	81	23 mil

Table 3.1: Model details

3.3 Real time detection

For real time emotion detection the first step is facial recognition. There are many algorithms that perform this task, conventional feature extraction methods, deep learning, and even transfer learning [8]. For real time facial recognition some criteria to consider for choosing an algorithm is speed in real time, low complexity and low computational expense.

One such algorithm satisfying these criteria is Haar cascades. This is an algorithm that detects objects using edge or line detection using Haar features [24]. A Haar feature is a specific calculation performed in adjacent rectangular regions in a detection window [14]. Haar cascades use a cascading window and the calculations are done in this region to detect an object. Some examples of Haar features for facial recognition are shown in figure 3.6.

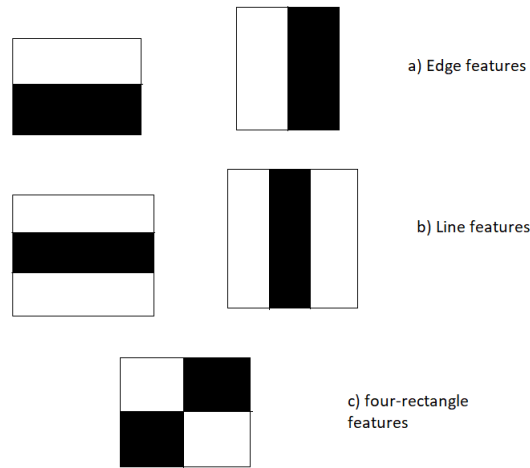


Figure 3.6: Caption

These features are used to detect faces based on the intensity of pixels in adjacent regions [6]. For example, the bridge of the nose is usually brighter compared to the region on either side of it. A Haar feature that detects lines can detect the nose by using this interesting property. In this manner using a cascading window the algorithm scans for features this way and tries to classify whether there is a face. However, this algorithm tends to predict more false positives [10]. For the purpose of this project it is not a concerning issue as a small fraction of false positives will not greatly impact the output.

Ethical Use of Data

The primary dataset used for this project is FER+ dataset which contains human participants. It is a publicly available dataset free to download from the respective GitHub repository. It is licensed under MIT license which allows any person obtaining the data to deal in the software/data without restriction, including without limitation to the rights to use, copy, modify, merge, publish, distribute, and sublicense.

The program must detect emotions in real time which raises the potential issue of using human participants to test and evaluate the program. To overcome this, instead of using human participants a publicly available and licensed dataset will be used for evaluation and I will be a test subject myself to evaluate real time emotion detection. This eliminates the need for human participants. I have read the ethical guidelines and will follow them.

Design

5.1 Model training

After the right dataset and the models are selected they have to be trained. The dataset chosen is FER+ which must first be pre-processed then analyzed to make it usable and well rounded to help the model generalize well. The pre-trained models chosen are VGG16, VGG19, InceptionV3 and Xception. Through some trial and error the number of layers to be unlocked and trained for each model have to be calculated, ideally at least three convolution layers must be unlocked.

The final layers consist of three dense layers with the last layer being a softmax layer. These models are then fine-tuned on the dataset then tested. State of the art models achieved around 73% [11] accuracy so it is reasonable to expect the models to achieve an accuracy around 55-65%.

Categorical cross entropy will be used as the loss function during training. The number of epoch each model will run for will be determined through trial and error so that the model doesn't over or under fit. The development of the models and training takes place primarily on paperspace platform which has access to cloud GPUs for faster processing and training.

5.2 Real time detection

The trained models are then imported on to a local machine. Visual Studio code is used to develop the software written in python. Real time emotion detection takes place in two steps, first is facial recognition and then emotion recognition from the detected face by passing it through the fine-tuned CNN model. Using OpenCV and haar classifiers facial detection will be done.

Haar classifiers detect faces based on line or edge detection. This method consists of features called Haar features. Haar cascade uses a cascading window and computes features in every window and tries to classify whether it is an object or not. For facial detection there are specific Haar features, they are stored in an XML file which can be imported via OpenCV and used for detection.

The detected face will be pre-processed to adjust for size and convert to black and white colour format then passed through the CNN. The obtained prediction will be displayed on screen.

5.3 Output

The final output will be a real time feed from the camera on to the display. A square bounding box will highlight the detected face and above the box the detected emotion along with confidence value will be displayed as shown in 5.1.

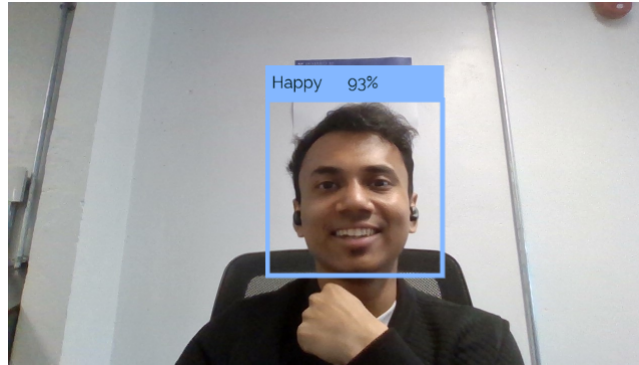


Figure 5.1: User Interface Mockup

5.4 Original design

In the original design AffectNet was the preferred dataset. It was ideal given the vastness, containing nearly 450,000 images with a resolution of 224x224 in RGB format. The subjects in the dataset are diverse, from a variety of different ethnic and racial backgrounds in varying lighting and environment conditions. However due to access issues with the dataset it could not be used.

The design was planned around this dataset so VGG16 and AlexNet were the initial pre-trained models selected. AlexNet was chosen due to strong feature extraction ability and it's ability to converge fast [17] Additionally only the three initial layers and last three dense layers of the networks were planned to be kept unlocked

5.5 Changes to original design

As the preferred dataset was not available the alternative dataset FER+ was used. This dataset is not as vast and diverse as AffectNet containing only 28,709 images and the images are black and white with a resolution of 48x48. The pre-trained models were trained on RGB images so working with black and white images may have affected the output.

AlexNet was replaced with InceptionV3 as AlexNet only accepts images larger than or equal to 224x224 resolution. Resizing the FER+ images to a resolution almost five times larger would not have yielded reliable outputs as the resized images would be blurred with additional noise. Two additional models were trained and tested as well. Which are VGG19 and Xception.

Instead of unlocking just the first three layers more layers were unlocked. The number of layers depended on how many layers were required to enclose the first three convolution layers. Through some trial and error of training on different number of unlocked layers around three to four convolution layers seemed to be the ideal number to improve accuracy.

Implementation

6.1 Data preprocessing

The software was developed in two stages. In the first stage the models were developed and fine-tuned on the paperspace platform using RTX4000 GPU for training related tasks. TensorFlow was the primary deep learning library used and FER+ was the chosen dataset. The images are stored in a CSV file so they must be converted from a CSV to numpy format then reshaped to be processed as an image. The number of samples are not evenly distributed (mainly disgust class) across all classes hence data augmentation techniques had to be implemented to even out distribution of samples. The distribution is shown below.

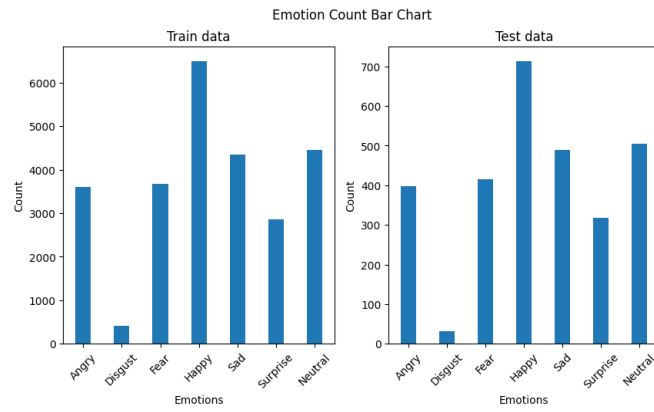


Figure 6.1: Data distribution before augmentation

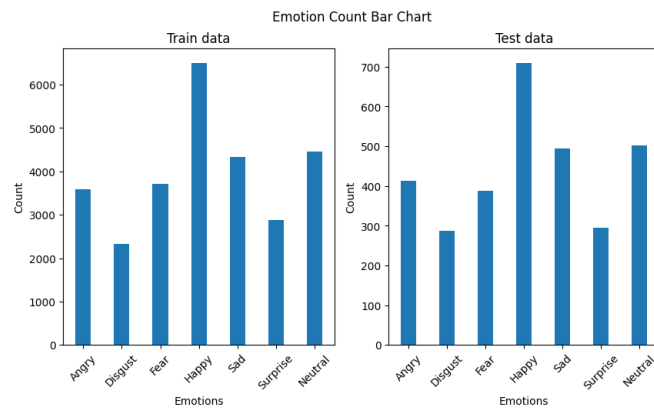


Figure 6.2: Data distribution after augmentation

6.2 Model fine-tuning

The pre-trained models chosen are VGG16, VGG19, Xception and InceptionV3. The models were downloaded and initially the first three layers were unlocked and trained on the FER+ dataset. However in later runs additional layers were unlocked to help increase accuracy and help feature extraction as only using three layers limited the models ability to learn.

The idea was to unlock at least three convolution layers for the networks. The pre-trained models were trained on RGB images and due to this reason they only accept three channel image inputs. To compensate for this the single channel images had the layer duplicated twice to make it compatible with the model. This operation was made part of the sequential network model.

The models were trained for 20 epochs and the training vs validation loss graphs were plotted. Using this information the appropriate number of epochs were determined based on convergence of the loss so the model doesn't over fit. The metrics table is shown below. The metrics graphs for training versus validation are as shown in appendix A

Model	Accuracy	F1 Score	Epochs	No. layers unlocked	Training Time(mins)
VGG16	61.2%	62.0%	12	10	3.2
VGG19	60.2%	60.4%	15	11	4.8
InceptionV3	50.8%	50.8%	15	13	5.0
Xception	59.7%	59.3%	8	15	4.9

Table 6.1: Metrics table

VGG16 achieved the highest accuracy on the test data. Closely followed by VGG19. Despite the number of parameters VGG16 has (138 million) it trained in relatively lesser epochs. InceptionV3 did not perform as well as expected even though it has a higher accuracy on the ImageNet dataset [22]. The VGG16 is closely followed by Xception which is an extreme version of InceptionV3.

6.3 Real time detection

The weights of the network and JSON file for model architecture were downloaded and tested on the local machine. All four models were tested on myself in real time using OpenCV and Haar classifiers on VS code on a local machine. The Haar cascades are an object detection algorithm for detecting faces using edge or line features. The detected face is pre-processed to adjust for size and convert to black and white colour format then passed through the CNN.

VGG16 performed the best, it was able to predict the displayed emotion and the output was stable. VGG19 performed quite well too but the output was not stable, it changed quite often. The other models could not accurately predict displayed emotions. Hence VGG16 was chosen for real time emotion detection. The output is shown in 6.3

The program access' laptop camera and displays the live feed on screen. In the camera feed there is a bounding box to highlight the detected face and the confidence value beside it. The output obtained from running the software is shown below.

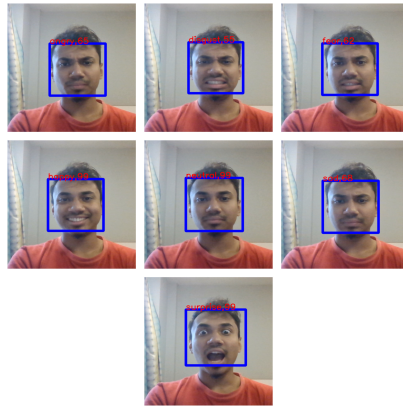


Figure 6.3: Real time emotion detection output

Evaluation

The output was achieved as intended with the model achieving 61% accuracy. It was able to predict displayed emotions in real time using myself as a subject. However, there were many changes in the implementation from the design proposal. They are listed below along with the strengths and weakness of the implemented project.

1. The software achieved real time emotion detection and was able to accomplish the primary aim of this project
2. The fine-tuned VGG16 model was chosen to be used for real time detection as it had the highest accuracy score and results were relatively stable in real time.
3. The model was able to do a good job a predicting emotions that had more samples in their respective class in the dataset.
4. The real time prediction had some difficulty separating sad and fear classes. Disgust class had very few samples, so data augmentation had to be applied to to increase the samples by five times. Due to data augmentation on the small number of samples the model predicts disgust in certain facial configurations better than the other.
5. The preffered dataset (AffectNet) was not granted access to so the back-up dataset (FER+) had to be used. Due to which model did not generalize well as the dataset images were smaller in resolution(48x48), black and white and much less in sample size(28,709).
6. As the back-up dataset was used, AlexNet pre-trained model had to be reconsidered as it did not work with the dataset. So InceptionV3 had to be used.
7. VGG19 and Xception were not considered initially but it was trained and tested as well. However, VGG16 performed better.
8. Due to the nature of the dataset only one emotional model could be tested (Ekman's Model) as opposed to the two models proposed.
9. Due to computational constraints the model could not be tested on CREMA-D dataset.
10. The number of layers unlocked in the models differ from the initial proposal. More layers were unlocked to improve feature extraction and get better accuracy.

Learning Points

This project gave me the opportunity to learn a lot about an up and coming field in computer science of real time emotion recognition. It was filled with challenges and required a lot of research and learning to solve them.

1. A significant amount of research was put into learning various emotional models and the challenges of facial expression detection.
2. Working with large datasets was relatively new to me, so working with numpy to manipulate and handle datasets was a new skill gained
3. Sometimes datasets have issues so I had to learn how to find, analyze and fix these issues in the pre-processing phase. Data visualisation techniques had to be applied to understand the data
4. One such issue was class imbalance. Data augmentation techniques were applied to help even the distribution
5. There were a lot of changes from the initially proposed design which were unexpected. Learning to expect such shortfalls and plan ahead in advance was an important skill learned
6. The proposed dataset was not granted access to so the alternative dataset was used. Due to this one of the proposed pre-trained models had to be replaced. Even though this was an issue faced this opened up the avenue to test more models and compare results
7. This was my first time working with fine-tuning pre-trained models. I had to learn how to import the models, unlock layers required for training and add additional layers for pre-processing the input images
8. Since this is the first time I worked with a large dataset and a computationally expensive program I realized the importance of resource and storage management. Processed data was saved and loaded when needed instead of being computed during each run
9. Real time emotion recognition was an interesting and challenging part of the project. I did not have much experience with computer vision and OpenCV libraries prior to this project. I learned about face detection algorithms and how to implement them using OpenCV

10. A lot of skills and knowledge were acquired during the course of this project ranging from handling pre-trained models to working with computer vision. The most important skill I believe I gained was risk management and dealing with problems that arise during project execution that have a high impact on the project. In the case of this project it was not getting access to the preferred dataset and figuring out how to resolve it in a smart way without compromising the quality of the project
11. In the future in order avoid pitfalls like this I would make sure to check access and availability of such resources required for the project during the research phase and make sure if it can be acquired
12. In case an unexpected problem arises and resources for a project can't be acquired, I would learn to plan better and be prepared for all contingencies.
13. Another thing I would like to do is not make the project too dependent on a single resource for future projects. With respect to this project, in the initial proposed design I built the project around the dataset. The selection of the pre-trained models were focused on the dataset without much room for error or change. Building around one resource creates a single point of failure which can adversely effect the project

Professional Issues

This is a project that involves development of an emotion recognition system using transfer and deep learning techniques. To see how this project relates to British Computer Society (BCS) Code of Conduct the various aspects of the code will be examined in regard to this project

9.1 Public Interest

1. This project aligns with the principle of public interest as it aims to create a system that can predict human emotions in real-time. Such technology can have many applications, including improving mental health support, human-computer interaction, education sectors to assess student engagement, and user experience in digital products.
2. All the data used and accessed for the purpose of this project is public data licensed under MIT license which allows any person obtaining the data to deal in the software/data without restriction, including without limitation to the rights to use, copy, modify, merge, publish, distribute, and sublicense.

9.2 Professional Competence and Integrity

1. All the work undertaken as a part of this project is within my range of professional competence as can be seen from the completion of work. The project also reflects the commitment to professional competence and developing professional knowledge, skills and competence by leveraging transfer learning techniques to achieve emotion recognition.
2. The relevant ethical guideline were followed and practiced for data collection, handling, storage and software testing. There were no human participants for testing except myself.

9.3 Duty to Relevant Authorities

1. Professional responsibilities in regard to project was carried out with due care and diligence under the oversight of my respective supervisors. Work was carried out under my supervisors requirements and they were consulted whenever there were project work related queries. I accept professional responsibility for all work done in the execution of this project
2. The project complies with the relevant laws regarding data privacy and and ethical use of human-supplied data. Necessary permissions were obtained to work with human data.

9.4 Duty to the Profession

1. This project contributes to the field of computer science by advancing the field of emotion recognition which has various applications. Sharing knowledge and findings from this project with the professional community can help promote best practices in the development of similar systems.

Conclusion

This research aimed to detect emotions in real time by leveraging transfer learning.

The primary aim of this project to identify human emotions with the help of transfer learning was achieved. The following objectives were also achieved.

1. Choosing the right pre-trained model: The models chosen are VGG16, VGG19, InceptionV3 and Xception. They were selected on the criteria mentioned in 3 which were briefly speed, accuracy, parameter size and nature of input.
2. Model should be generalized: The preferred dataset (AffecNet) was not granted access to so the alternate dataset (FER+) had to be used instead which effected the models ability to generalize well but it still managed to perform well.

Many different models with varying architectures were trained and tested. Some interesting findings in this project were as follows.

1. Simpler networks performed better with respect to accuracy and convergence which can be seen from the graphs A.
2. During real time testing the effect of data distribution and augmentation could be seen in the output. The network easily identifies emotions which had the most samples in their respective class and some difficulty predicting data augmented classes and classes with lesser samples.

It was demonstrated in this project that the models that performed better had simpler architectures and lesser layers. It was also shown that the models ability to generalize and be fine-tuned well depended on the nature of the dataset. ImageNet models fine-tuned on RGB images with larger resolutions in comparison to FER+ would have performed better as these models were initially trained on images of a similar format.

The results of this paper show that leveraging transfer learning is an effective method to achieve real time emotion recognition using lower computational power, resources, training time and datasets.

Bibliography

- [1] Ashi Agarwal and Seba Susan. “Emotion Recognition from Masked Faces using Inception-v3”. In: Mar. 3, 2023, pp. 1–6. DOI: 10.1109/RAIT57693.2023.10126777.
- [2] M. a. H. Akhand et al. “Facial Emotion Recognition Using Transfer Learning in the Deep CNN”. In: *Electronics* 10.9 (Jan. 2021). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 1036. ISSN: 2079-9292. DOI: 10.3390/electronics10091036. URL: <https://www.mdpi.com/2079-9292/10/9/1036> (visited on 06/23/2023).
- [3] Aqeel Anwar. *Difference between AlexNet, VGGNet, ResNet and Inception*. Medium. Jan. 22, 2022. URL: <https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaecccc96> (visited on 06/26/2023).
- [4] Emad Barsoum et al. *Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution*. Sept. 23, 2016. DOI: 10.48550/arXiv.1608.01041. arXiv: 1608.01041[cs]. URL: <http://arxiv.org/abs/1608.01041> (visited on 06/29/2023).
- [5] François Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. Apr. 4, 2017. DOI: 10.48550/arXiv.1610.02357. arXiv: 1610.02357[cs]. URL: <http://arxiv.org/abs/1610.02357> (visited on 09/13/2023).
- [6] Li Cuimei et al. “Human face detection algorithm via Haar cascade classifier combined with three additional classifiers”. In: *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*. 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI). Oct. 2017, pp. 483–487. DOI: 10.1109/ICEMI.2017.8265863.
- [7] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition. ISSN: 1063-6919. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [8] *Face Recognition using Transfer learning on a pre-trained model(VGG16)*. URL: <https://www.linkedin.com/pulse/face-recognition-using-transfer-learning-pre-trained-mayukh-borana> (visited on 08/21/2023).
- [9] Divya Garg and Gyanendra K. Verma. “Emotion Recognition in Valence-Arousal Space from Multi-channel EEG data and Wavelet based Deep Learning Framework”. In: *Procedia Computer Science*. Third International Conference on Computing and Network Communications (Co-CoNet’19) 171 (Jan. 1, 2020), pp. 857–867. ISSN: 1877-0509. DOI: 10.

- 1016/j.procs.2020.04.093. URL: <https://www.sciencedirect.com/science/article/pii/S1877050920310644> (visited on 06/26/2023).
- [10] Abhishek Jaiswal. *Guide to Haar Cascade Algorithm with Object Detection Example*. Analytics Vidhya. Apr. 1, 2022. URL: <https://www.analyticsvidhya.com/blog/2022/04/object-detection-using-haar-cascade-opencv/> (visited on 09/20/2023).
 - [11] Yousif Khairuddin and Zhuofa Chen. *Facial Emotion Recognition: State of the Art Performance on FER2013*. May 8, 2021. DOI: 10.48550/arXiv.2105.03588. arXiv: 2105.03588[cs]. URL: <http://arxiv.org/abs/2105.03588> (visited on 06/29/2023).
 - [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 07/06/2023).
 - [13] Siyang Li et al. “Facial Expression Recognition In-the-Wild with Deep Pre-trained Models”. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 181–190. ISBN: 978-3-031-25075-0. DOI: 10.1007/978-3-031-25075-0_14.
 - [14] Aditya Mittal. *Haar Cascades, Explained*. Analytics Vidhya. June 26, 2021. URL: <https://medium.com/analytics-vidhya/haar-cascades-explained-38210e57970d> (visited on 09/19/2023).
 - [15] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild”. In: *IEEE Transactions on Affective Computing* 10.1 (Jan. 1, 2019), pp. 18–31. ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2017.2740923. arXiv: 1708.03985[cs]. URL: <http://arxiv.org/abs/1708.03985> (visited on 06/29/2023).
 - [16] Hong-Wei Ng et al. “Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI ’15. New York, NY, USA: Association for Computing Machinery, Nov. 9, 2015, pp. 443–449. ISBN: 978-1-4503-3912-4. DOI: 10.1145/2818346.2830593. URL: <https://dl.acm.org/doi/10.1145/2818346.2830593> (visited on 06/25/2023).
 - [17] Sarmela A-P Raja Sekaran, Chin Poo Lee, and Kian Ming Lim. “Facial Emotion Recognition Using Transfer Learning of AlexNet”. In: *2021 9th International Conference on Information and Communication Technology (ICoICT)*. 2021 9th International Conference on Information and Communication Technology (ICoICT). Aug. 2021, pp. 170–174. DOI: 10.1109/ICoICT52021.2021.9527512.
 - [18] Ricardo Ribani and Mauricio Marengoni. “A Survey of Transfer Learning for Convolutional Neural Networks”. In: *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*. 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T). ISSN: 2474-0705. Oct. 2019, pp. 47–57. DOI: 10.1109/SIBGRAPI-T.2019.00010.

- [19] Shamoil Shaees et al. “Facial Emotion Recognition Using Transfer Learning”. In: *2020 International Conference on Computing and Information Technology (ICCIT-1441)*. 2020 International Conference on Computing and Information Technology (ICCIT-1441). Sept. 2020, pp. 1–5. DOI: 10.1109/ICCIT-144147971.2020.9213757.
- [20] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Apr. 10, 2015. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556[cs]. URL: <http://arxiv.org/abs/1409.1556> (visited on 07/06/2023).
- [21] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. version: 3. Dec. 11, 2015. DOI: 10.48550/arXiv.1512.00567. arXiv: 1512.00567[cs]. URL: <http://arxiv.org/abs/1512.00567> (visited on 09/01/2023).
- [22] Keras Team. *Keras documentation: Keras Applications*. URL: <https://keras.io/api/applications/> (visited on 08/31/2023).
- [23] Sik-Ho Tsang. *Review: Xception — With Depthwise Separable Convolution, Better Than Inception-v3 (Image...* Medium. Mar. 20, 2019. URL: <https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568> (visited on 09/12/2023).
- [24] P. Viola and M. Jones. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. ISSN: 1063-6919. Dec. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.

Training Metrics

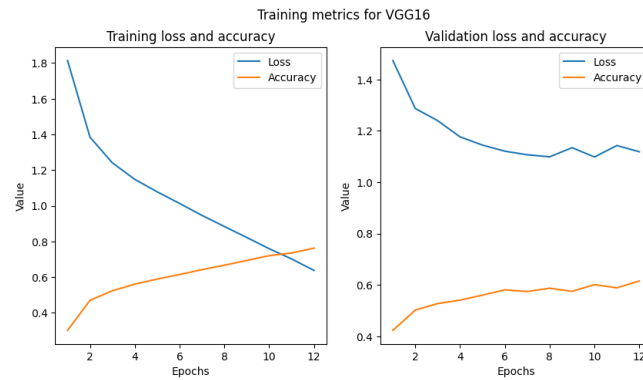


Figure A.1: VGG16 Metrics Graph



Figure A.2: VGG19 Metrics Graph

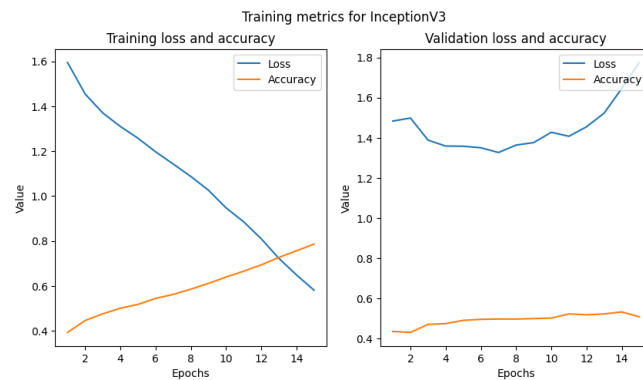


Figure A.3: InceptionV3 Metrics Graph

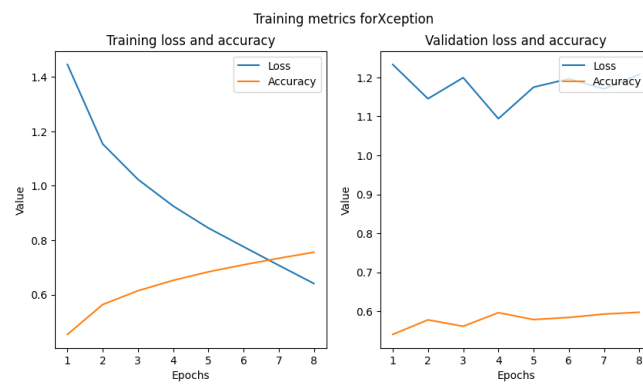


Figure A.4: Xception Metrics Graph