Multimodal Sentiment Analysis: Enhanced Methods for Sentiment Exploration

Taha Juzer Husain (20231431)
College of Science and Engineering
National University of Ireland, Galway

Supervisors
Dr. Josephine Griffith

In partial fulfillment of the requirements for the degree of
MSc in Computer Science (Data Analytics)
August 31, 2021

DECLARATION

I, Taha Juzer Husain, do hereby declare that this thesis entitled Multimodal Sentiment Analysis: Enhanced Methods for Sentiment Exploration is a bonafide record of research work done by me for the award of MSc in Computer Science (Data Analytics) from National University of Ireland, Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: _____

Acknowledgement

I would like to thank my research supervisor, Dr. Josephine Griffith, who has been an excellent research mentor to me throughout this research work. Thank you very much for your constant supervision, guidance, and patience which has helped me to reach the destination. I am grateful to you for your availability to listen and support me for my idea, implementation and technical discussions.

Furthermore, my deepest gratitude to my family and friends for supporting me throughout my journey in completing my Masters.

Abstract

This research studies different techniques for video and image analysis for sentiment prediction. As there is an increase in sharing views on social media, people use different methods or modals to share views such as text, images, and GIFs. Analyzing all the modals together to predict the sentiment can give better results. Different models for video analysis using deep learning methods such as Conv_LSTM, Conv_3D have been implemented. These models take into consideration the spatial and temporal features in a video. For image classification, two models i.e. Xception and MobileNet_V2 are implemented using the transfer learning approach. Multiple datasets are used for image and video analysis. The dataset used for video classification is provided by MIT. For image classification, the dataset used is from Cornell University. After conducting different experiments on the models, results are compared to examine which model provides the best performance. The accuracy of models is calculated by comparing the predicted values to the ground truth. The final models from image and video with the highest accuracies are used for sentiment prediction. A multimodal framework is proposed to use the different modals and predict the sentiment.

Keywords: Mutimodal Sentiment Analysis, Convolution, 3D CNN, LSTM, Deep Learning

Contents

1 Introduction **1**

1.1 Research Questions 2

1.2 Research Methodology 2

1.3 Organization of the research 3

2 Background and Related Work **4**

2.1 Introduction 4

2.2 Textual Sentiment Analysis 4

2.2.1 Machine Learning models 4

2.2.2 Deep Learning models 5

2.3 Visual Sentiment Analysis 6

3 Dataset Generation and Analysis **12**

3.1 Introduction 12

3.2 Dataset Generation for GIFs 12

3.3 Dataset Analysis for GIFs 13

3.4 Dataset Analysis for Images 14

3.5 Dataset Analysis for Text 16

4 Methodologies and Models **17**

4.1 Introduction 17

4.2 Conv_LSTM 17
4.3 3D CNN 19
4.4 Xception Model 19
4.5 MobileNet_V2 Model 20
4.6 Text Model 21
4.7 Training Steps for models 23

5 Experimental Analysis and Results **24**
5.1 Introduction 24
5.2 Image Modal 24
5.2.1 Xception Model 25
5.2.2 MobileNet_V2 Model 27
5.2.3 Comparison of results from image models 28
5.3 Text Modal 30
5.4 GIF Modal 30
5.4.1 Conv_LSTM model 30
5.4.2 Conv_3D model 32
5.4.3 Conv_3D model with Transfer Learning 36
5.4.4 Comparison of results for GIF models 38
5.5 Multimodal Framework 38

6 Conclusion and Future Work **40**
6.1 Overall Discussion 40
6.2 Conclusion 41
6.3 Future Work 42

References **46**

A Code **47**

List of Figures

2.1 Architecture of CNN model 8

3.1 Example of Positive image 15
3.2 Example of negative image 15

4.1 Inner structure of Conv_LSTM 18
4.2 Transforming of 2D image into 3D tensor 18
4.3 Kernel Sliding on 3D Data 20
4.4 Architecture of Xception model 21
4.5 Architecture of Mobilenet_V2 model 22

5.1 Accuracy graph of Xception model 26
5.2 Loss values of Xception model 26
5.3 Accuracy graph of MobileNet_V2 model 28
5.4 Loss values of MobileNet_V2 model 28
5.5 Accuracy comparison of image models 29
5.6 Architecture of Conv_LSTM model 31
5.7 Accuracy Graph for Conv_LSTM model 32
5.8 Loss Graph for Conv_LSTM model 33
5.9 Architecture of Conv_3D model 34
5.10 Accuracy Graph for Conv_3D model 35

5.11 Loss Graph for Conv_3D model 35
5.12 Accuracy Graph for Conv_3D model with transfer learning 37
5.13 Loss Graph for Conv_3D model with transfer learning 37

List of Tables

5.1 Accuracy comparison for Image models 29
5.2 Accuracy comparison for GIF models 38

Chapter 1

Introduction

As the use of social media has increased exponentially, users leverage multiple modals such as text, images, and videos to share their views. Studies have suggested that there is a correlation between the sentiment of online users and real-world scenarios. The analysis of sentiment in this content can provide information on a variety of topics, including politics, e-commerce, and predicting box office receipts for movies. As per the study by Yuan et al. (2013), people are more likely to share animated GIF files than static photographs, and the total proportion of visual content on social media was 36% in 2013 and has been steadily increasing since then.

Because of its capacity to display dynamic content, tell tales, and convey emotions, the Graphics Interchange Format (GIF) has become the internet's favorite file type (Gygli and Soleymani, 2016). Moreover, there are advantages of using GIFs such as small size, silence without sound which make them easy to consume as compared to long videos (Bakhshi et al., 2016). According to recent studies, Tumblr generates more than 23 million GIFs every day, Sina or Weibo has over half a billion short GIFs, and more than 71 percent of web articles feature short GIFs (Liu et al., 2020). As more and more users use GIFs in their posts

on social media, discovering the knowledge embedded in multimedia becomes a task of great significance. Unimodal sentiment analysis takes into consideration a single modal such as text or image for sentiment prediction. However, there could be mixed sentiments in the modals which could be missed. Taking different modalities into account can give a more holistic result in the task of sentiment prediction. This research aims to implement different models for GIF and image analysis, and compare the results from the models.

1.1 Research Questions

1. Can the use of different deep learning models such as CNN, 3D CNN, Conv_LSTM, help in sentiment prediction when using a GIF?

2. Can the analysis of spatiotemporal features instead of only spatial features improve the accuracy of GIF models for sentiment analysis?

3. Can the use of transfer learning and pretrained models improve the accuracy of sentiment prediction for GIF and images?

1.2 Research Methodology

There are many techniques implemented on image and text modals for sentiment analysis. This research aims to build different models for GIF, and image analysis, and compare the performance of models.

GIFs are small videos which are widely being used on social media to convey messages. There are multiple things which needs to be taken into consideration while working with GIFs. This research finds different ways to analyse GIF videos. There are different features in the images, in the frame transitions, and the sequence of frames. Thus, we need to consider the spatiotemporal features

while working with short videos. There are multiple models created and tested on the GIFs. Different set of features are used in training and evaluation. Different techniques such as CNN, 3D CNN, Conv_LSTM are tested to see which model yields better results. Moreover, use of transfer learning and pretrained models such as MobileNet_V2 and Xception, are done to see if better results can be achieved using the image modal.

1.3 Organization of the research

In the following chapters, different models and modals have been discussed. The order of the discussion is as follows:

- Chapter 2 presents the relevant background research and the related work.
- Chapter 3 describes the datasets used in the research.
- Chapter 4 presents details of all the methods and models used in the research.
- Chapter 5 discusses all the experiments and results from the different models and modals.
- Chapter 6 summarises the research discussing conclusions and future work.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter presents a brief discussion regarding different methods in text, image, and GIFs such as machine learning, and deep learning for sentiment prediction.

2.2 Textual Sentiment Analysis

There are different models built to perform textual sentiment analysis by training the model with a dataset. These models include machine learning models and deep learning models.

2.2.1 Machine Learning models

In machine learning models, a dataset is used to train the model and predict the sentiment of a given sentence based on the training. The steps included in training the models include gathering the data, dataset preprocessing such as tokenization, removal of stop words, and removal of non-English words. Later,

feature extraction is performed such as bigram, trigrams, and Parts of Speech (POS) tagging. Post feature extraction, feature selection is performed using NLP, statistical-based, clustering-based, and hybrid-based methods. Different machine learning algorithms can be used to train the model, in Mandloi and Patel (2020) three algorithms are used. In the Naive Bayes algorithm, the Bayes theorem is used for classification problems. It is one of the most simple and effective algorithms while working with texts. The second is Support Vector Machines (SVM), where classification is performed by drawing a hyperplane that differentiates between the classes which are plotted in n-dimensional space (Tammina, 2019). Lastly, the maximum entropy method is used for performing sentiment analysis. In this approach, the algorithm does not presume that the events are independent. It selects the data which is the best fit for the training data and has the maximum entropy among them.

Good performance can be achieved when using machine learning models, but a major problem is faced while performing the feature extraction and selection process. To overcome this problem, deep learning methods such as Recurrent Neural Networks (RNN) are used.

2.2.2 Deep Learning models

To overcome the problem of the feature extraction and selection process, deep learning models such as RNN are used. RNN helps in the sequential processing of textual data. While working with long texts, the RNN models face issues such as vanishing and exploding gradients. To overcome this problem, algorithms such as LSTM and GRU are used.

In Zarzour et al. (2021), LSTM and GRU models have been implemented for textual sentiment analysis. The model architecture consists of an embedding layer followed by LSTM/GRU layer followed by fully connected layers and a sentiment

prediction or an output layer. The embedding layer converts the string data into a numeric data vector, which is a sparse encoding as input and transforms into dense encoding data. The output of this layer is continuous data. LSTM not only tries to remember a part of the input but also tries to remember the context. GRU is a variation of LSTM which has fewer parameters and is simpler to calculate.

Results of experiments performed in Zarzour et al. (2021), show GRU outperforms LSTM and SVM models on training and testing sets. Training accuracy of 92% is achieved using GRU whereas the SVM model achieved less than 80% accuracy.

2.3 Visual Sentiment Analysis

While working with visuals, different features can help in classification. In Jou et al. (2014), the author discusses different feature representations of an image. Different feature representations can be used to learn the sentiment in the image. Four picture feature representations are employed with their past use or link to emotion recognition in visual content to explore how characteristics contribute in the prediction of viewer perceived emotion. The four representations are as follows:

- Color histograms - In HSV color space, a frame-level color histogram is produced for use in vision and affect computing.
- Face expression - A convolutional neural network with top-level one-vs-all SVMs with squared hinge losses is used to determine the face expression. The OpenCV library's Haar-like cascade is used to detect faces, and it uses facial expression recognition on the largest face as a feature for a 6-D vector of SVM score outputs.

- Image-based aesthetics: GIF frames are divided into 3x3 cells for aesthetic purposes, from which cell-level data such as the dark channel, luminance, sharpness, symmetry, white balance, colorfulness, color harmony, and eye sensitivity are computed.

- SentiBank: It consists of 1200 trained visual idea detectors that provide a mid-level representation of sentiment, as well as accompanying training photos from Flickr and a benchmark of 603 photo tweets covering a wide variety of 21 topics (Borth et al., 2013).

Learning features from the data is a difficult task. To improve the feature selection, techniques of deep learning can be used. Deep Learning tries to automatically select the best features by ensemble methods. While working with image data, CNN plays an important role.

CNN (Convolutional Neural Network) a class of deep neural networks has become dominant in working with computer vision tasks. CNN was first developed in the 1980s for the task of digit recognition. Multiple layers, such as convolution layers, pooling layers, and fully linked layers, make up a convolutional neural network. They use a backpropagation technique to learn spatial hierarchies of features automatically and adaptively. The first two layers of convolution and pooling layers perform the task of feature extraction while the later connected layers map the features to the output such as classification. A convolution layer is an important part of CNN; it is made up of mathematical processes like convolution, which is a form of linear operation. Pixel values are kept in a two-dimensional (2D) array in digital images, and at each image point, a small array of parameters called the kernel, which functions as a feature extractor, is applied. This makes CNNs highly efficient for image processing since a feature can occur anywhere in the given image. As all the layers are connected, extracted features can hierarchically and progressively become more complex. The parameters such

as kernels are improved during the training process, and training is carried out to minimize the difference between outputs and ground truth labels using an optimization technique such as backpropagation or gradient descent, among others. Figure 2.1 shows the CNN architecture [1].

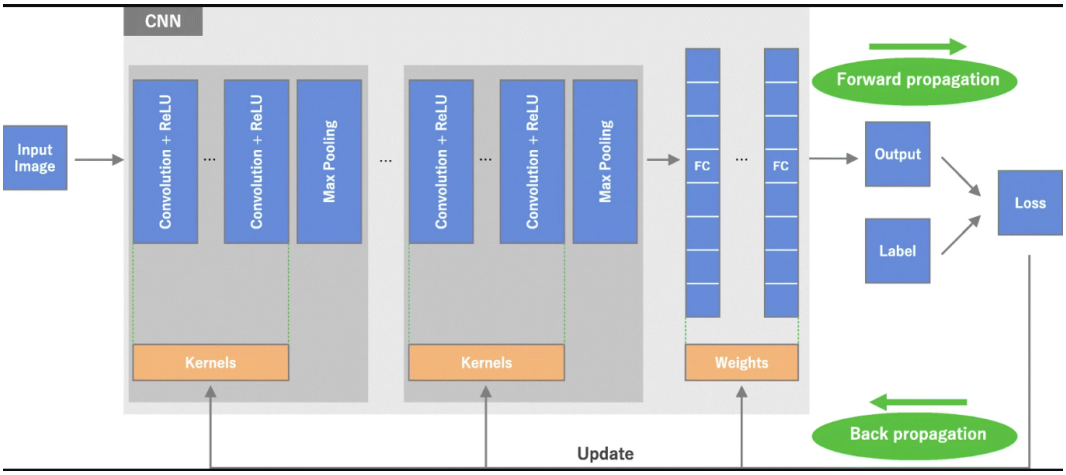

Figure 2.1: Architecture of CNN model

In Shirzad et al. (2020), three separate modules are created for text, images, and GIFs. Depending on the type of media in the tweet, they are passed to different modules in the framework. For the text, the model uses the VADER python library to analyze the polarity. To work with the images, a fine-tuned CNN model is used to calculate the sentiment score. To improve the performance of the CNN model, a method of transfer learning is used. Transfer learning is a method of reusing a pre-trained model for another task. We can use the weights of the initial layers of the model and train them further on a dataset (Tammina, 2019). While working with images, lower level features can be similar, so the initially trained layers of the model can be used to learn the basic features. Later, by adding fully connected layers we can fine tune the model for the required task. Transfer learning can be used for classification, regression, and clustering

[1] https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9

problems. A pretrained model of VGG 16 is used in the image module. If a tweet contains a GIF file, the file is divided into frames and is passed to the image module. Finally, the scores are combined for all the models with the ratio of 2,1,1 for text, image, and GIF to provide an average sentiment score. The score calculated is a float number between -1 for extreme negative and +1 for extreme positive.

While using CNN on images, it tries to extract all the features from the image. In this process, important features can be missed. It should be noted, in the case of facial expressions, much of the clues come from fewer parts of the image, for example, mouth or eyes. Other body parts such as ears or hairs do not contribute to facial expression recognition. To focus on the important parts of the face, an attention mechanism is implemented (Minaee and Abdolrashidi, 2019). For better performance, generally more layers are added to the network. In Minaee and Abdolrashidi (2019), an approach based on the attention mechanism is implemented which focuses on feature-rich parts of the face and outperforms the recent works in the field. The network described in Minaee and Abdolrashidi (2019) uses fewer layers as compared to state-of-art methods. To concentrate on the key features, an extra mechanism is used through the spatial transformer network to focus on face regions. The spatial transformer (the localization network) consists of two convolution layers (each followed by max-pooling and ReLU), and two fully connected layers. The results from the paper show that with a smaller network and attention mechanism, better results in facial expression recognition were achieved.

In the previous research, full utilization of heterogeneous data was not performed. To improve on the attention model, the Multi-Attention model was proposed in Kim and Lee (2020). A Multi-Attention Recurrent Neural Network (MA-RNN) is used to perform multimodal sentiment analysis. In this proposed

network, there are two attention layers and a Bidirectional Gated Recurrent Neural Network (BiGRU). The first attention layer is used for data fusion and reduction of dimensionality. The second layer is used for the augmentation of BiGRU to capture the key parts of the contextual information. Using the sequence of information in a video for understanding the inter-utterance relationships and finding contextual evidence is important for sentiment classification. For sequence data, Long Short- Term Memory (LSTM) and Gated Recurrent Unit (GRU) is tried. By performing empirical evaluations, no significant difference was found in LSTM and GRU models. However, by using the GRU network, a 6.2% improvement was found as compared to LSTM.

In another research, LSTM was used with an attention mechanism to learn the sentiment from images. In the proposed system in You et al. (2015), it uses LSTM with an attention mechanism to learn the sentiment from an image. Initially, the system builds a semantic tree structure based on sentence parsing and tries to align textual words and image regions for analysis. Next, the system tries to learn a joint text-visual semantic representation by using an attention mechanism with LSTM and auxiliary semantic learning tasks. Tree-structured LSTM is used for joint textual-visual sentiment analysis which helps in better mapping of textual words and image regions. A bilinear attention model is used for the feature representation and to prevent a single modality from dominating the model. Moreover, a Siamese network is adopted as an additional task to learn the semantic embedding between the text and image. This paper helps in working with the images but lacks methods for the GIF and video analysis.

In Chen and Picard (2016), the author proposes the use of a 3D CNN model for GIF analyses. Substantial portions of GIFs are made of animation and anime which makes facial expression recognition difficult. While dividing the GIFs into frames, important temporal information related to motion is not considered. To

address this problem, 3D CNN is implemented so spatiotemporal information instead of spatial features are extracted. The model was evaluated on the GIFGIF dataset by MIT and Normalized mean squared error (NMSE) of 0.6652 ± 0.0545 was achieved.

In summary of the research papers discussed above, most of the work done on textual and visual sentiment analysis is based on deep learning methods. For textual sentiment analysis, good results were achieved using deep learning methods such as GRUs and LSTMs (Zarzour et al., 2021). Machine learning models have been also implemented for textual sentiment analysis, however, the problem of feature extraction and selection decreases the efficiency of models. While working with visuals, many different techniques have been implemented and promising results are achieved. Many of the methods involve Deep Learning methods such as CNN, and LSTMs. CNN helps in extracting features from images. Once the features are extracted, they can be used to train the model. Improvements on CNN are implemented such as 3D CNN which uses the spatiotemporal features for training (Chen and Picard, 2016). In another paper, LSTM with attention was implemented to predict the sentiment from images (You et al., 2015). It should be noted that there were no papers that implemented the methods in conjunction. In the following research, I try to implement models using multiple methods to understand if better results could be achieved in the task of sentiment prediction.

Chapter 3

Dataset Generation and Analysis

3.1 Introduction

In this chapter, datasets for text, image, and GIFs are discussed. An overview of the datasets are presented, e.g., statistics of the dataset like frames per video and analysis of the datasets in order to understand how the data will be used in the models.

3.2 Dataset Generation for GIFs

The GIF dataset is provided by MIT [1]. It is a public dataset which contains GIFs for 17 different emotions. MIT provides a friendly RESTful API to download the GIFs. Parameters for the API include:

- mID: It acts as an identifier for the category. Multiple mIDs can be passed in square brackets i.e. [ ] for multiple categories.
- Limit: Number of GIFs to be fetched in a single API call.

[1] `http://gifgif.media.mit.edu/`

- Key: Public key used to fetch the data.

The API call returns a JSON file that includes the date added, and the link for the GIF file.

GIFs of four categories i.e. 'Happiness', 'Surprise', 'Angry', and 'Sad' are downloaded from the website. These are further divided into positive (Happiness, and Surprise), and negative (Angry, and Sad) categories.

3.3 Dataset Analysis for GIFs

Statistics of frames of GIFs for the positive category is as follows:

- Count: Total count of GIFs is 1000.
- Mean: The mean number of frames in the positive category is 28 frames.
- Min: Minimum number of frames is 2 frames.
- Max: Maximum number of frames is 257 frames.

25% of GIFs contains 12 frames whereas 75% of GIFs contains 34 frames in the positive category.

Statistics of frames of GIFs for the negative category is as follows:

- Count: Total count of GIFs is 1000.
- Mean: The mean number of frames in the negative category is 29 frames.
- Min: Minimum number of frames is 2 frames.
- Max: Maximum number of frames is 303 frames.

In the negative category, 25% of GIFs contains 13 frames whereas 75% of GIFs contains 35 frames.

This shows an equal distribution of frames in each category.

Consecutive frames in GIFs tend to be similar with no or minimal changes. To reduce the data size and to consider the maximum number of frames in GIF, frames in multiples of four are considered, for eg, frames 0, 4, 8 ... . To make the data consistent for models, all videos are padded with dummy frames if the count of the frame is less than 16. The final dimensions of the GIF dataset are (16, 112, 112, 3), where there are 16 frames and each frame is 112 x 112 pixels with 3 channels.

3.4 Dataset Analysis for Images

The image dataset from Cornell university consists of different emotions such as joy, anger, disgust, fear, etc (Peng et al., 2015). The categories are divided into positive, and negative categories. Emotions such as Joy and Surprise are clubbed into the positive category. Anger and Sadness emotions are considered into the negative category. Fig 3.1 and fig 3.2 gives an example of both the categories.

Data augmentation is performed on the dataset where different transformations such as shear transformation, zoom, and horizontal flip are done. The dataset is divided into train and test set for model evaluation. There are 1000 images in the training set and 320 images in the test set. The images are resized according to the model requirements.

The Shape of the dataset for the MobileNet_V2 model is (224, 224, 3), whereas, the shape of the dataset for the Xception model is (299, 299, 3).

Figure 3.1: Example of Positive image

Figure 3.2: Example of negative image

3.5 Dataset Analysis for Text

For the text modal, the dataset used is provided by Vadicamo et al. (2017). The dataset consists of 1179957 tweets. Each tweet contains the score for different sentiments such as positive, negative, and neutral. The highest score defines the category of the tweet. For the purpose of testing the model, a test set of 600 tweets is made from the original dataset. 200 tweets from each category i.e. 'Positive', 'Negative', and 'Neutral' is taken to balance the test set. No pre-processing is performed on the test set, as the VADER python library handles them in the analyzer.

Chapter 4

Methodologies and Models

4.1 Introduction

In this chapter, the different methods which are used for the GIF and image analysis are explained. The different methods implemented for GIF analysis are Conv_LSTM and Conv_3D. A transfer learning approach is used for the image analysis where different models such as MobileNet_V2, and Xception from Keras are fine-tuned and used.

4.2 Conv_LSTM

LSTMs are used while working with sequential data. LSTMs remember long-term dependencies in the network. They use gates to control what information is remembered. Videos can also be considered as sequential data of frames. While using LSTMs for video analysis important information such as spatiotemporal can be missed. They handle input-to-state and state-to-state transitions in which no spatial information is encoded. To overcome this problem, Conv_LSTM is used where outputs of cells are 3D tensors whose last two dimensions are spatial (rows

and columns) (Xingjian et al., 2015). It is similar to the LSTM layer but the input transformations and recurrent transformations are both convolutional in this layer. Figure 4.1 represents the inner structure of Conv_LSTM (Xingjian et al., 2015). Figure 4.2 represents the transformation of 2D image into 3D tensor (Xingjian et al., 2015).

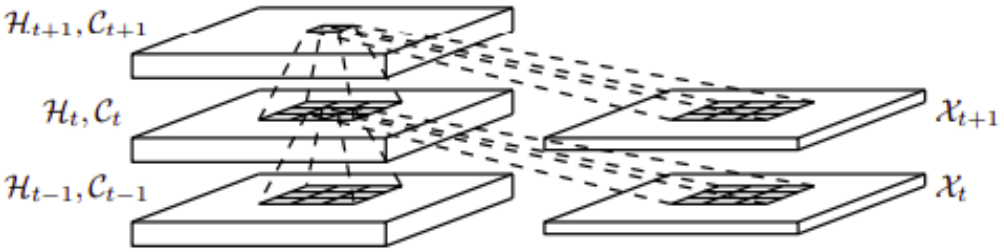

Figure 4.1: Inner structure of Conv_LSTM

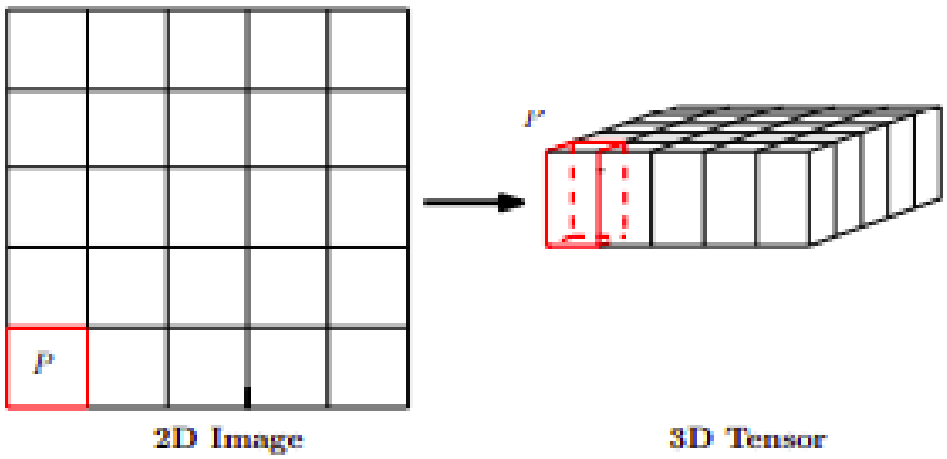

Figure 4.2: Transforming of 2D image into 3D tensor

4.3 3D CNN

CNN models are used to learn features from images. Each input image is passed through a series of convolutional layers with filters (kernels), pooling, fully connected layers and softmax is applied in the last layer for classification.

There are different types of CNN such as 1D, 2D, and 3D CNN. In 2D CNN's, convolutions are applied to the 2D feature maps for learning features from the spatial dimensions only. While working with the video analysis, motion information is encoded in the sequence of frames. 3D CNN can be used to compute features from spatial and temporal dimensions. 3D convolution is achieved by convolving 3D kernels to the cube which is formed by using multiple frames together. Using this architecture, feature maps are connected to multiple contagious frames in the previous layer which helps in capturing motion information (Li et al., 2017). Fig 4.3 represents kernel sliding on 3D data [1].

Another 3D CNN model is trained using a transfer learning approach (Tran et al., 2015). Pretrained weights of the model are used for the initial layers. Later layers are trained on the data for classification. Sports1M model has been trained earlier on sports 1M dataset for action recognition. The model is further trained on the GIF dataset.

4.4 Xception Model

The Xception model was developed by Google in 2017. It is a CNN model which is based completely on depth-wise separable convolution layers. The idea behind this model was to decouple the mapping of cross-channel correlations and the spatial correlations in the feature maps of CNN can be entirely decoupled. The

[1] https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610

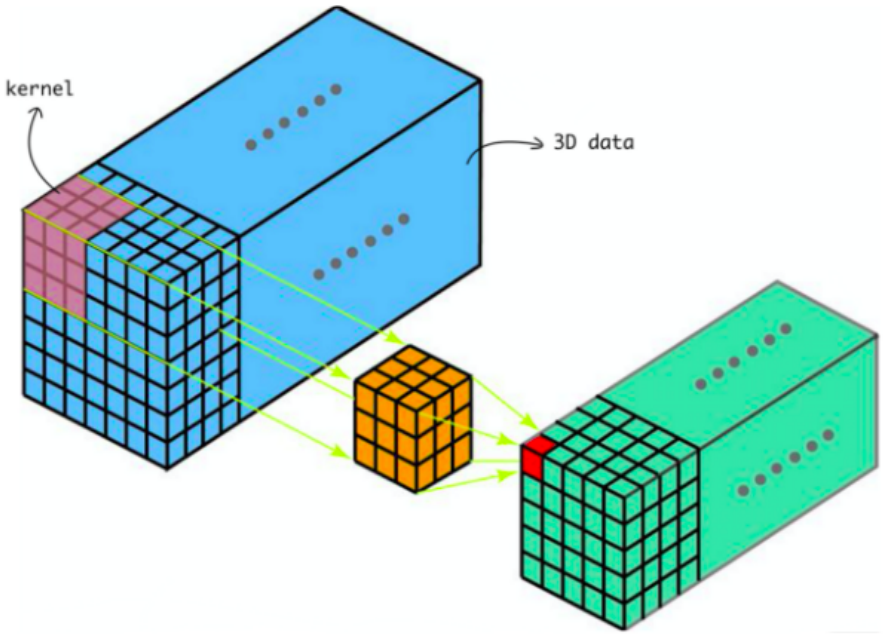

Figure 4.3: Kernel Sliding on 3D Data

model consists of 36 convolution layers which form the basis of feature extraction. The input shape of the model is (299, 299, 3). Fig 4.4 represents the architecture of the Xception model (Chollet, 2017). Keras provides a direct implementation of the Xception model, which can be used for transfer learning. The weights of feature extraction layers can be set to non-trainable. The last layers of the model can be trained to perform classification.

4.5 MobileNet_V2 Model

The MobileNet_v2 model was introduced by Google in 2019. It is an improvement over the MobileNet model which can be used for classification, object detection, and semantic segmentation. It is a lightweight model with fewer parameters to train. It contains an initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers. The kernel size used in the model is 3x3 and dropouts

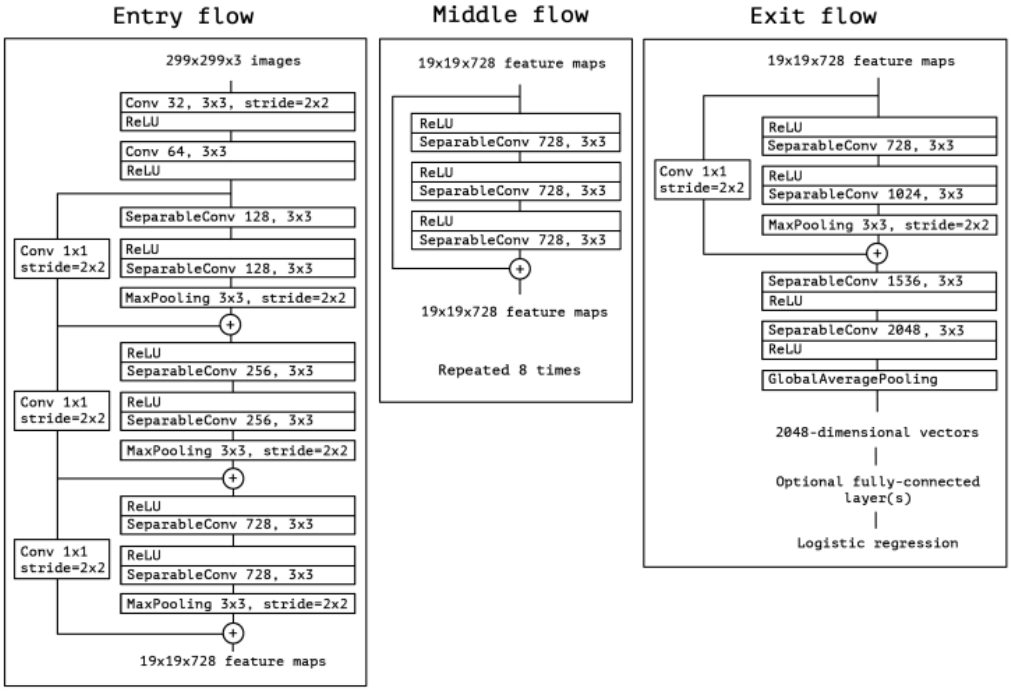

Figure 4.4: Architecture of Xception model

and batch normalization is used during the training of model (Sandler et al., 2019). The major features of V2 over V1 are 1) The linear bottlenecks between different layers, and 2) The shortcut connections between the bottlenecks. Fig 4.5 represents the architecture of the MobileNet_V2 model [1]. Transfer learning is performed using this model. Weights of initial layers are set as nontrainable and later layers are trained on the data.

4.6 Text Model

For textual sentiment analysis, VADER python library is used. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a lexicon and rule based sentiment analysis tool which is specially trained for sentiments expressed on social media. It

[1] https://ai.googleblog.com/2018/04/mobilenetv2-next-generation-of-on.html

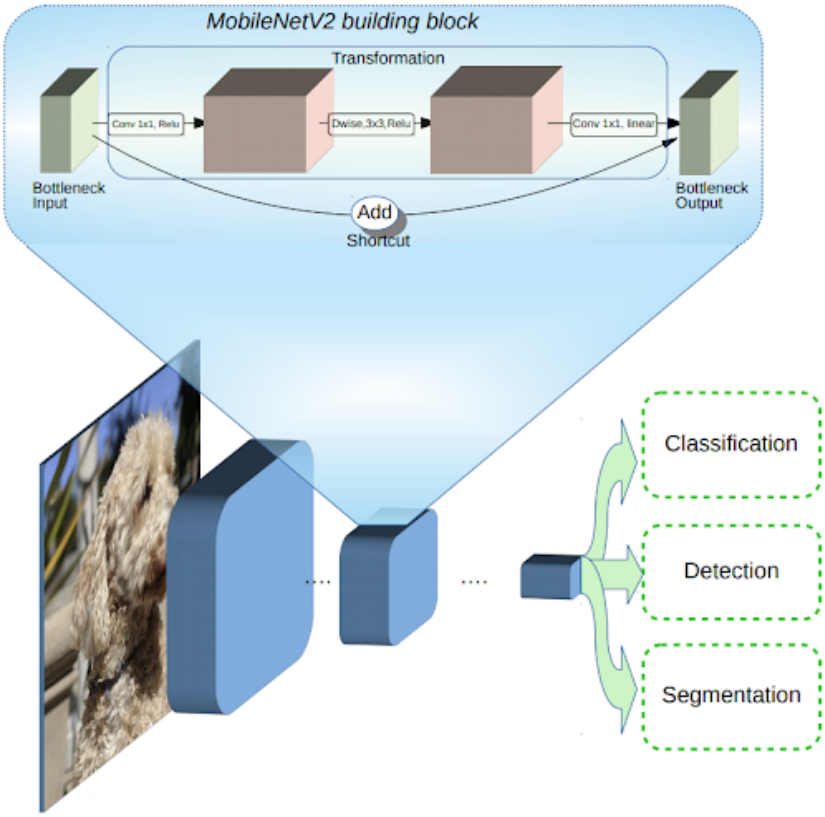

Figure 4.5: Architecture of Mobilenet_V2 model

uses a combination of qualitative and quantitative methods to produce sentiment lexicon for predicting sentiment of text as positive or negative (Hutto and Gilbert, 2014).

The analyzer method returns a compound score for the sentence. It is a metric that is scaled between -1 and 1. -1 represents extreme negative score, +1 represents extreme positive score. Different range of scores are as follows:

- positive: compound score >= 0.5
- neutral: (compound > -0.05) and (compound < 0.05)
- negative: compound <= -0.05

4.7 Training Steps for models

The general steps involved in training the above models are defined as follows:

Step 1: Data Preparation

All the models accept input in specified dimensions. The images need to be reshaped as per the requirements of the model. To resize the images, the OpenCV module is used. In the process of transfer learning, models expect predefined dimensions with which they were earlier trained. For video classification models, frame length also needs to be defined as per model.

Step 2: Model Definition

Different layers need to be defined in the model. If a transfer learning approach is being used, a predefined model can be imported which has its definition. Further layers are added to train the model on specific data. Other parameters such as loss function, optimizer, epochs, and batch size are defined. Keras is used to define the hidden layers and compiling and fitting the model.

Step 3: Model Training and Evaluation

The model is trained on the training data. Later the model is evaluated on the testing data. The accuracy of the model is calculated by finding how many correct predictions are performed by the model.

Chapter 5

Experimental Analysis and Results

5.1 Introduction

In this chapter, different experiments and results are discussed. There are three different models built for the GIF classification task. Experiments have been performed on Conv_LSTM, Conv_3D, and Conv_3D with a transfer learning approach. For image classification, two different models have been experimented with to see which model provides better results.

5.2 Image Modal

For the Image models, the dataset is divided into 75% training set and 25% testing set. The training set contains 1000 images of both the categories and the test set contains 320 images of both categories. The ImageDataGenerator from Keras is used to augment the data, resize the images, and create the training and testing set. The images are rescaled between 0 to 1.

5.2.1 Xception Model

Xception model is imported from Keras application library. The model parameters are as follows:

- Weights: Pretrained weights of the imagenet have been used to train the model.
- Optimizer: RMSProp optimizer has been used to train the model with a learning rate of 0.01.
- Loss: Sparse Categorical Crossentropy is used as the loss function for the model.
- Epochs: Model is trained for 20 epochs.

Additional layers of Global Average Pooling and dense layer with 2 neurons have been added for classification. The model is trained using the training set and evaluated based on the test set.

Fig 5.1 and fig 5.2 represent the accuracy and loss values of the trained model. The accuracy graph shows that the accuracy increases over each epoch and at final epoch accuracy of 86% is achieved. The loss graph also shows a downfall from first epoch. This indicates the model is learning the features.

Results of the Model:

- Training Accuracy : 86.4%
- Testing Accuracy : 79.3%

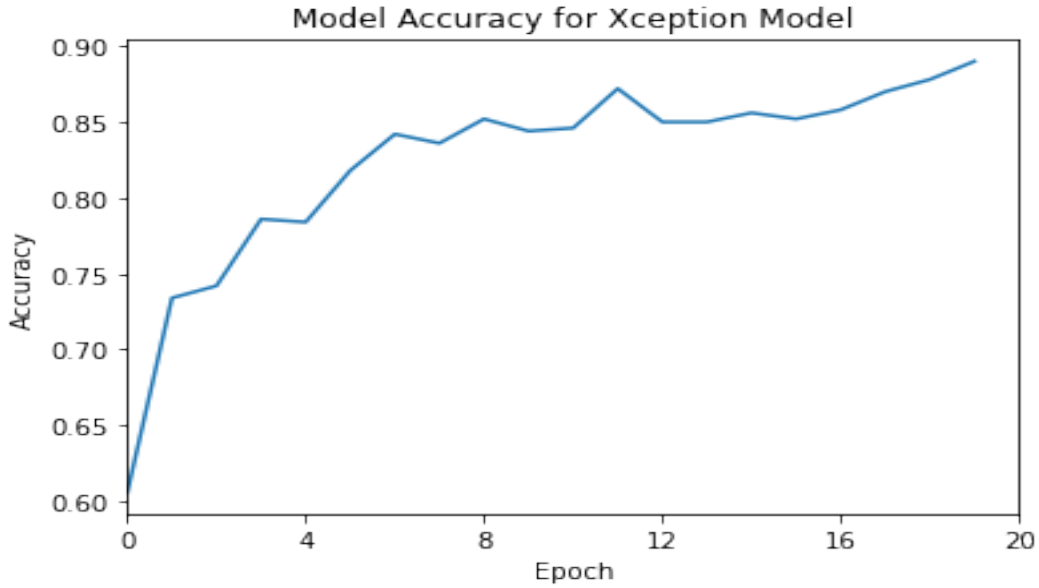

Figure 5.1: Accuracy graph of Xception model

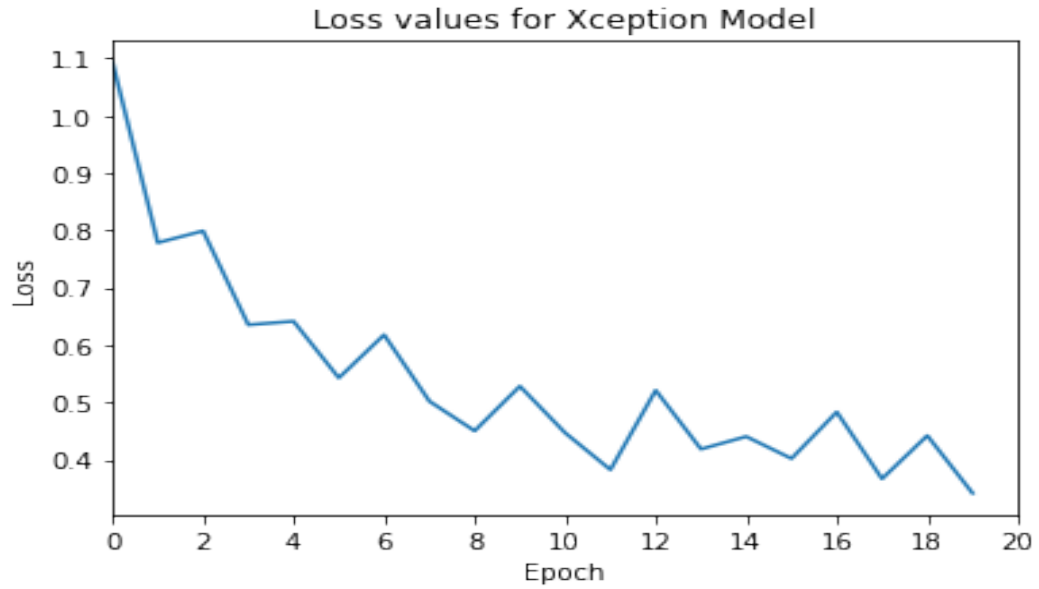

Figure 5.2: Loss values of Xception model

5.2.2 MobileNet_V2 Model

The model is imported from the Keras application library. The model parameters are as follows:

- Weights: Pretrained weights of the imagenet have been used to train the model.
- Optimizer: RMSProp optimizer has been used to train the model.
- Loss: Sparse Categorical Crossentropy is used as the loss function for the model.
- Epochs: Model is trained for 20 epochs with 80 steps per epoch.

Additional layers of Global Average Pooling and dense layer with 2 neurons have been added for classification. The model is trained using the training set and evaluated based on the test set.

Fig 5.3 and fig 5.4 represent the accuracy and loss values for the MobileNet_V2 model. Initially, the model training was slow and accuracy was increasing slowly. After epoch 14, there was a large increase in the accuracy of the model. Final accuracy at the last epoch was 87%. The Loss graph also shows a downfall from the start of training. There are rises but overall loss is decreasing over time. This indicates the model is reducing loss and learning features over the epochs.

Results of the Model:

- Training Accuracy : 86.9%
- Testing Accuracy : 75.6%

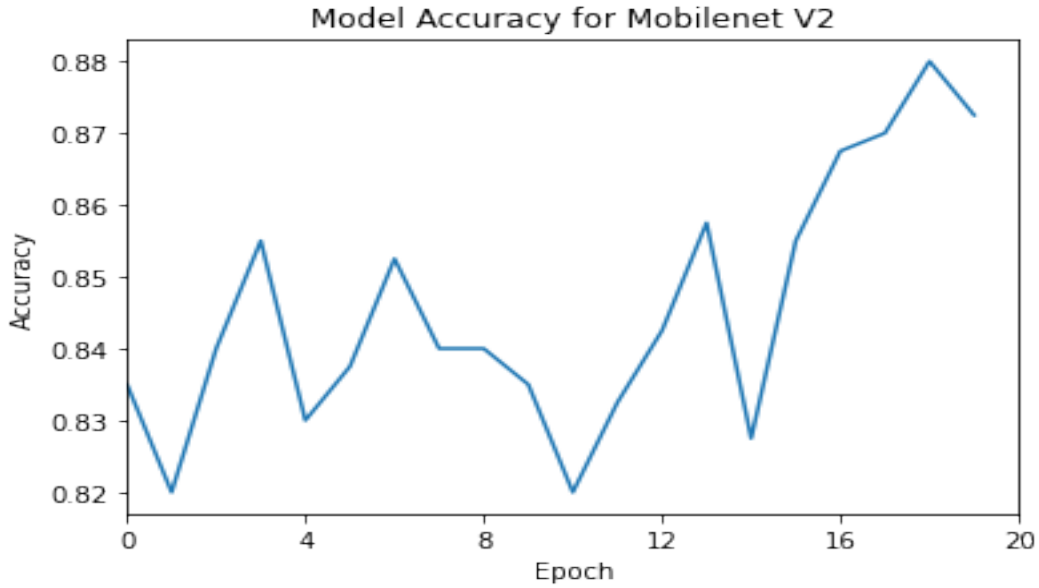

Figure 5.3: Accuracy graph of MobileNet_V2 model

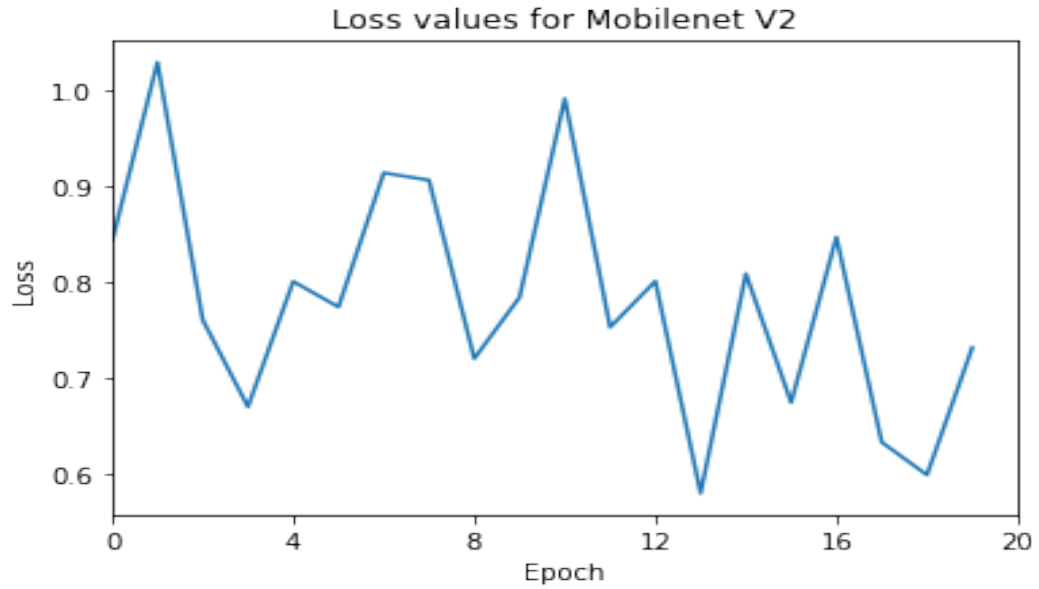

Figure 5.4: Loss values of MobileNet_V2 model

5.2.3 Comparison of results from image models

Figure 5.5 represents the comparison of accuracy for both models.

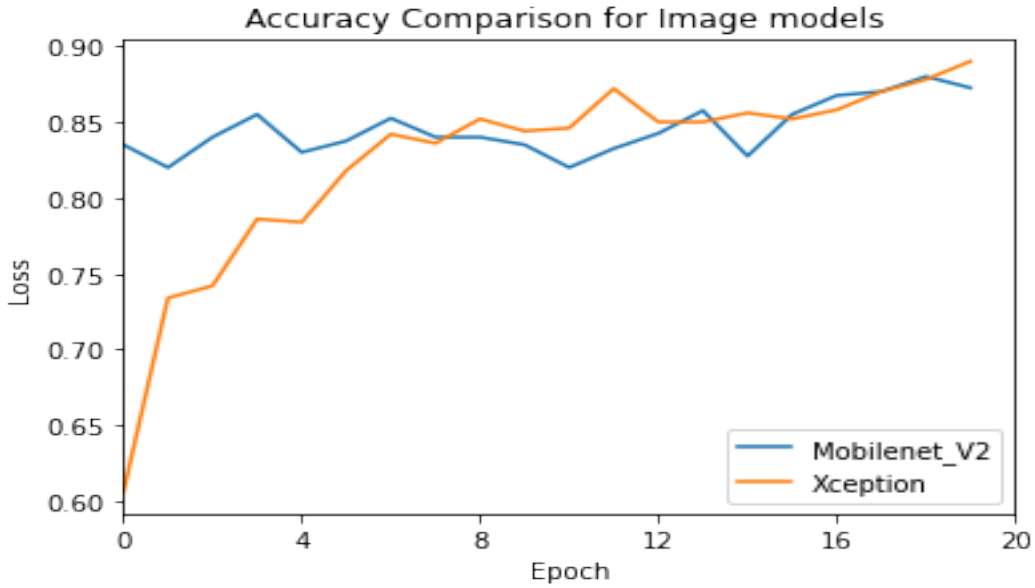

Figure 5.5: Accuracy comparison of image models

Model	Accuracy
VGG16 model by Shirzad et al. (2020)	83%
Xception	79.2%
MobileNet$_V$2	75.6%

Table 5.1: Accuracy comparison for Image models

As per table 5.1 it can be observed that the Xception model performs better on the training dataset whereas the MobileNet_V2 model performs better on the test dataset. As a transfer learning approach is being used in training the images, the models performs efficiently after little training on the dataset. The models use pretrained weights for initial layers. The difference between the scores of both the models is not very significant and the performance of both models can be considered equal.

5.3 Text Modal

Using the VADER python library, an accuracy of 74% was achieved on the test dataset. The predicted results are compared with the ground truth available. The analyzer performs decently on the data. Moreover, it also considers emoticons used in the tweets. Thus, it can be used without performing much preprocessing on textual tweet data.

5.4 GIF Modal

To prepare the dataset for GIF analysis, data is divided into 80-20% training and testing sets. OpenCV library is used to convert frames to NumPy array. There are 16 frames in each video and each frame is reshaped to 112 x 112 with three channels. Data is rescaled between 0 to 1.

There are three models built to analyze the GIFs. They are as follows:

5.4.1 Conv_LSTM model

All layers used in building the model are imported from Keras layers. The architecture of the model is shown in fig 5.6.

The model parameters used are as follows:

- Optimizer : Adam optimizer
- Loss : Sparse Categorical Entropy
- Epochs : 10 epochs
- Batch_size : 5
- Validation_split : 0.20

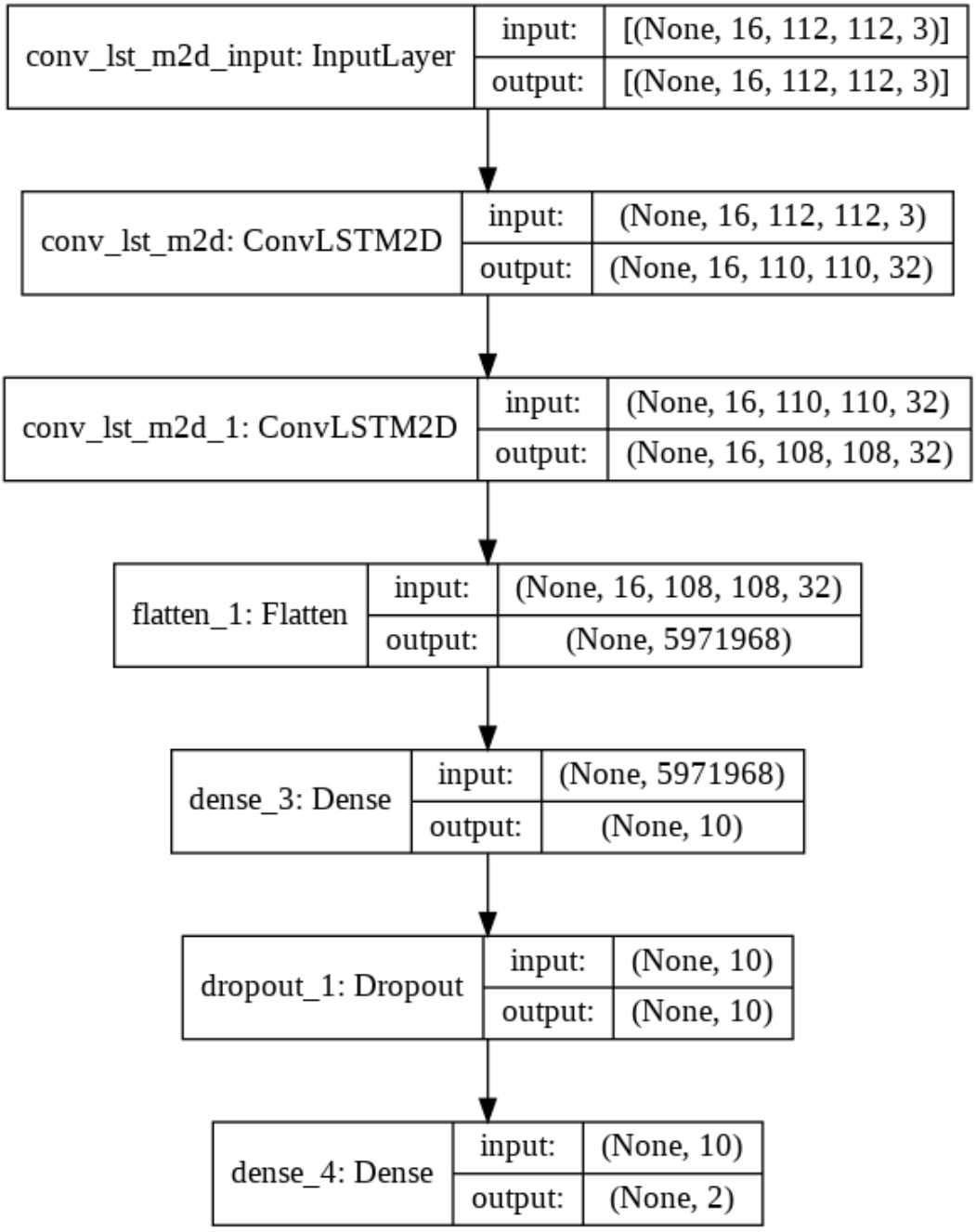

Figure 5.6: Architecture of Conv_LSTM model

Different experiments are performed on the model such as adding and removing layers, changing optimizer and changing loss function to binary cross-entropy.

Moreover, changing number of epochs had minimal or no changes on the results. Best results are achieved with the above model parameters.

Figure 5.7 and Figure 5.8 represent the accuracy and loss graphs for the trained model. The accuracy graph represents the training accuracy over 10 epochs. The accuracy increases in some epochs whereas decreases in the next epochs. The model is unable to learn the features as the epochs increases. The validation accuracy remains constant over the training, that shows the model is not performing well on the validation set. The loss values shows the model is learning but the loss is decreasing very slowly. The validation loss remains constant through out the training.

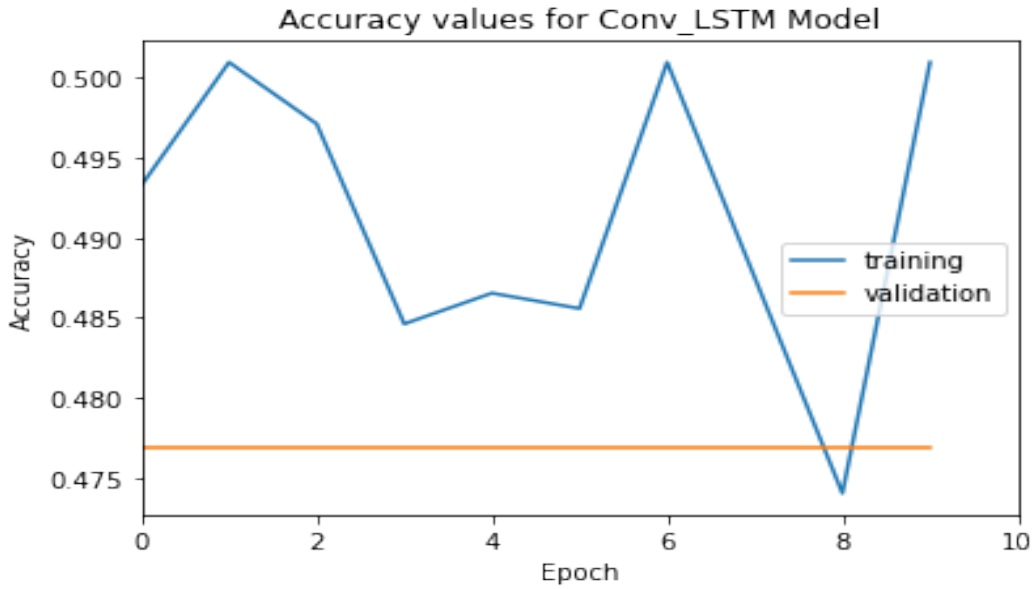

Figure 5.7: Accuracy Graph for Conv_LSTM model

5.4.2 Conv_3D model

Layers used in building the model are from Keras layers. The architecture of the model shown in fig 5.9 is taken from Tran et al. (2015). The architecture has been

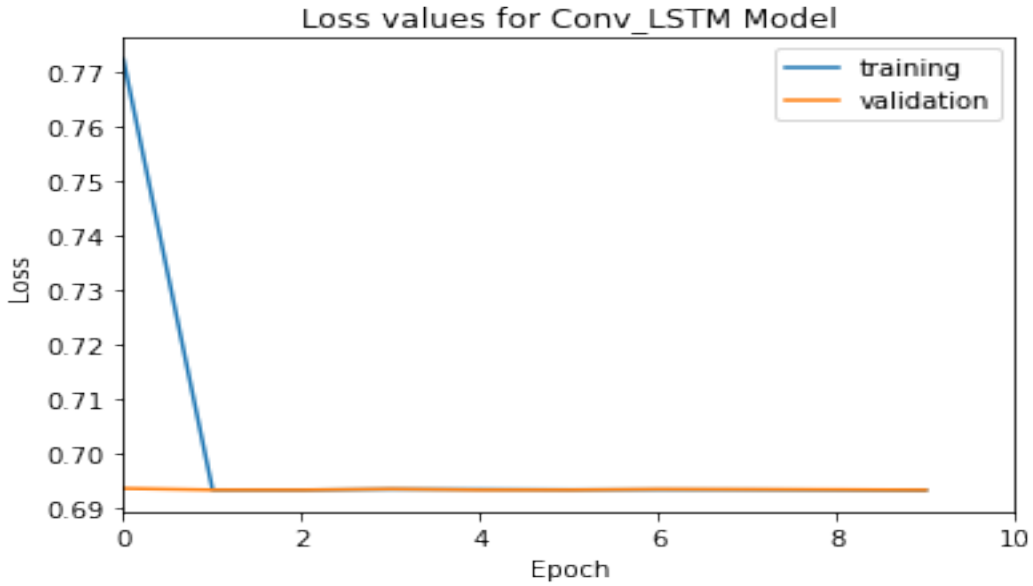

Figure 5.8: Loss Graph for Conv_LSTM model

modified with some layers. The original architecture had maxpooling3D layers however better results were achieved with averagepooling3D layers.

Model parameters are similar to the Conv_LSTM model. The Number of epochs used in training the model is 30 and the batch size is 100. Similar experiments were performed on the Conv_3D model to improve results. Best results are achieved with the current model.

Fig 5.10 and fig 5.11 represent the accuracy and loss graphs for the trained model. The accuracy graph represents the training and validation progress of the model. The training accuracy fluctuates, increasing in some epochs and decreasing in some. Validation accuracy also increases and decreases at some steps. Overall graphs shows the model is learning in the training. The loss graph also shows that the loss is decreasing with the epochs. The training and validation accuracy decreasing represents that the model is learning.

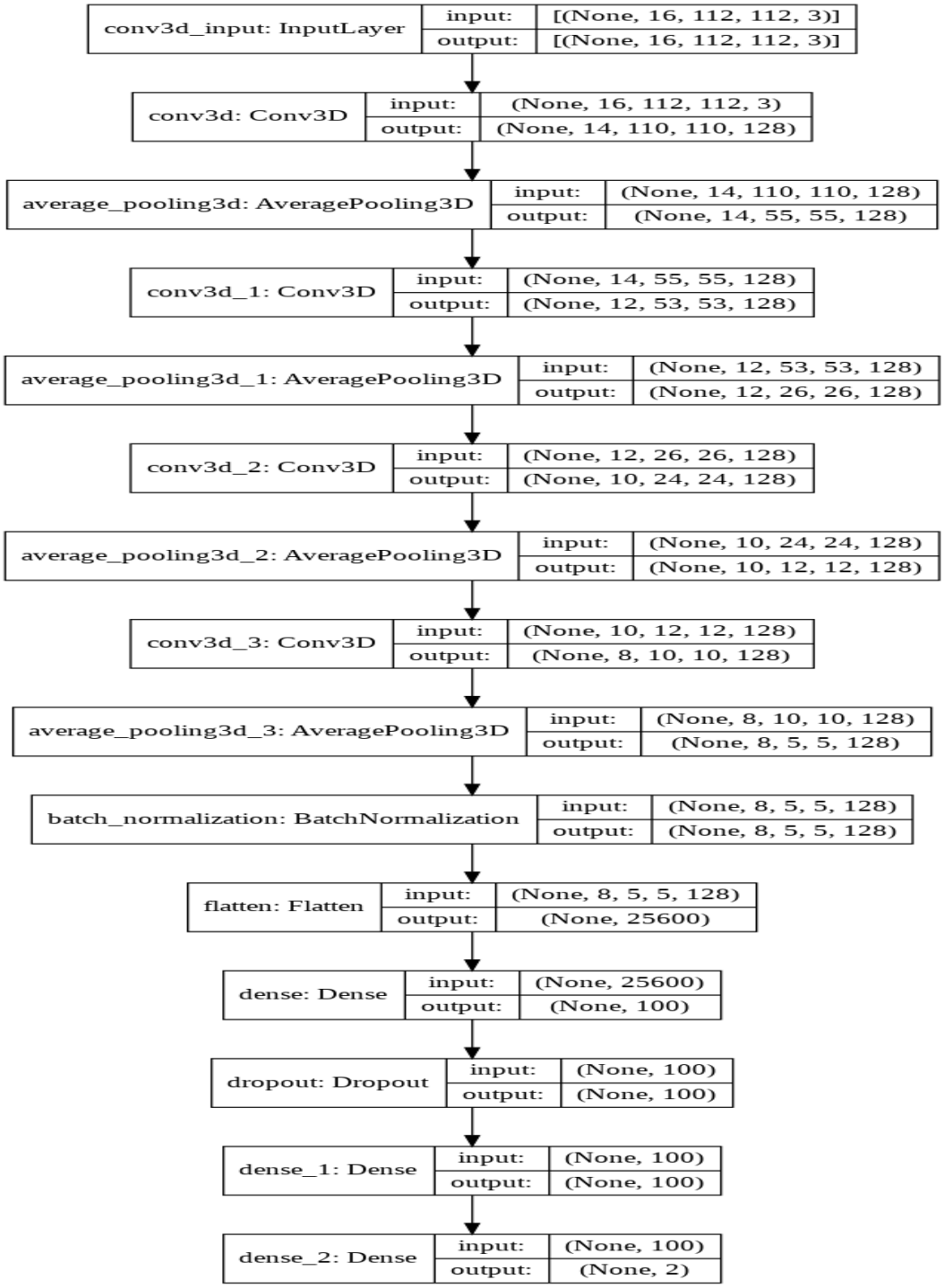

Figure 5.9: Architecture of Conv_3D model

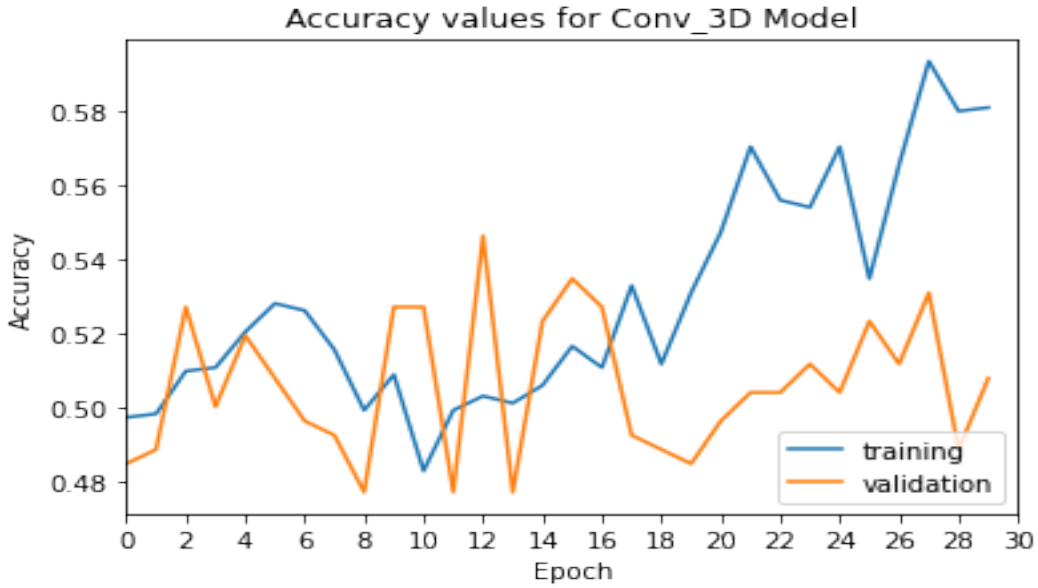

Figure 5.10: Accuracy Graph for Conv_3D model

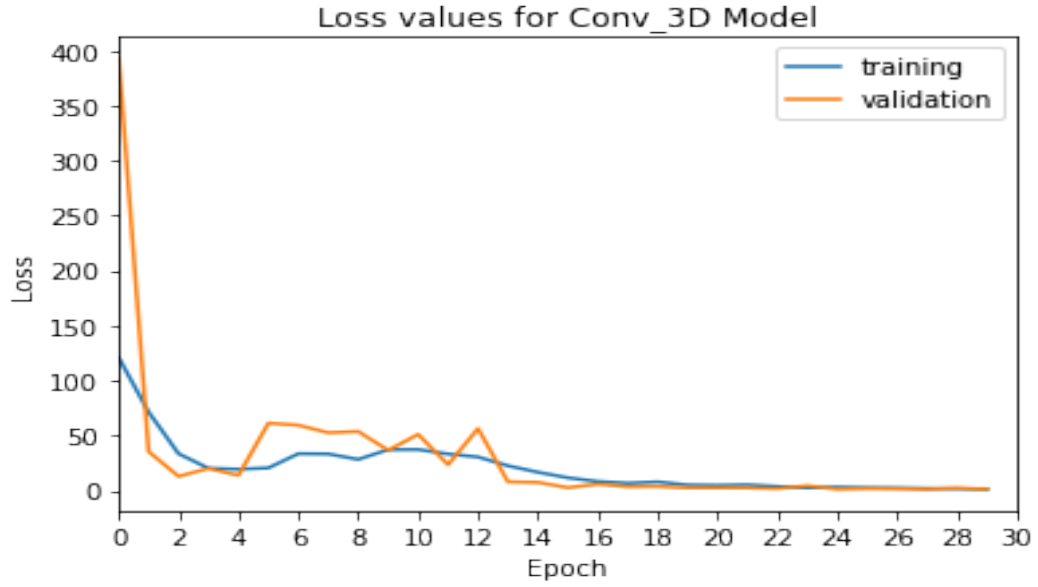

Figure 5.11: Loss Graph for Conv_3D model

5.4.3 Conv_3D model with Transfer Learning

The architecture and weights for the model are taken from Tran et al. (2015). The model is trained on the Sports1M dataset which contains videos of actions that are classified into different categories. Weights for initial Conv_3D layers are kept unchanged as they act as feature detectors. More layers are added to the model so it could learn the features of the current dataset. Layers added to the model are as follows:

- Convolution3D: Convolution3D layer is added with 512 nodes and a kernel of (3, 3, 3). The activation function used is leakyrelu.
- ZeroPadding3D : ZeroPadding3D layer with padding as (0. 1, 1).
- MaxPooling3D: Layer with pool size as (2, 2, 2) and strides as (2, 2, 2).
- Dense: Two dense layers with nodes 50 and 2 are added for classification.

Other parameters include loss as BinaryCrossEntropy, optimizer as Adam with learning rate as 0.1. The model is trained on 10 epochs and a batch size of 20. Higher number of epochs such as 20, and 30, had no significant changes on the results. Validation split of 0.20 is done on the training data.

Fig 5.12 and fig 5.13 represent the accuracy and loss graphs for the trained model. The accuracy graph shows the model is performing good in some epochs but it is unable to learn enough features to correctly predict in the next epochs. The model does not perform well on validation data as there is no increase in the accuracy. This overall represents that the model fails to train on the training set. The loss graph shows that the training loss is decreasing with the epochs, however, validation loss is constant in the whole training.

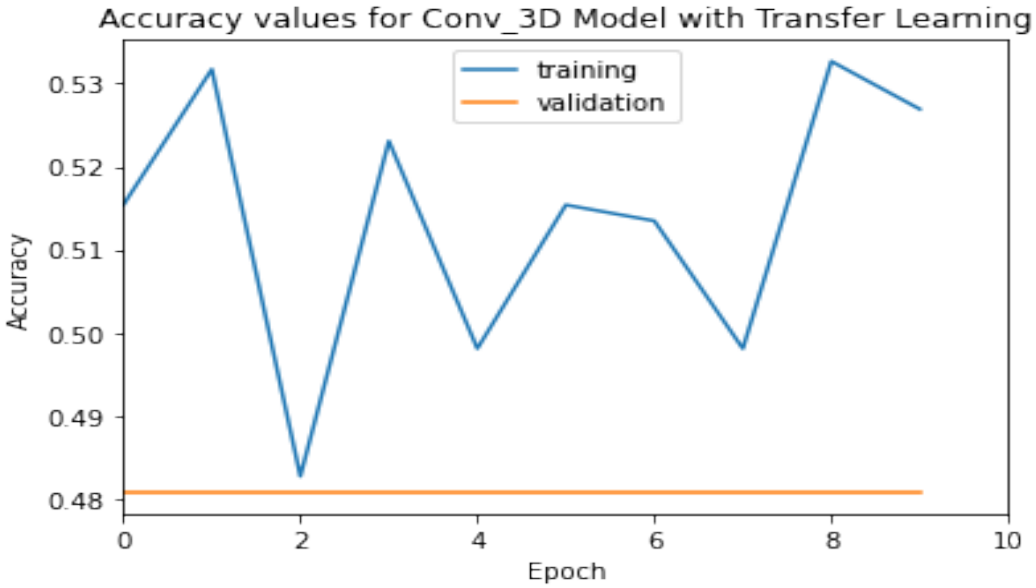

Figure 5.12: Accuracy Graph for Conv_3D model with transfer learning

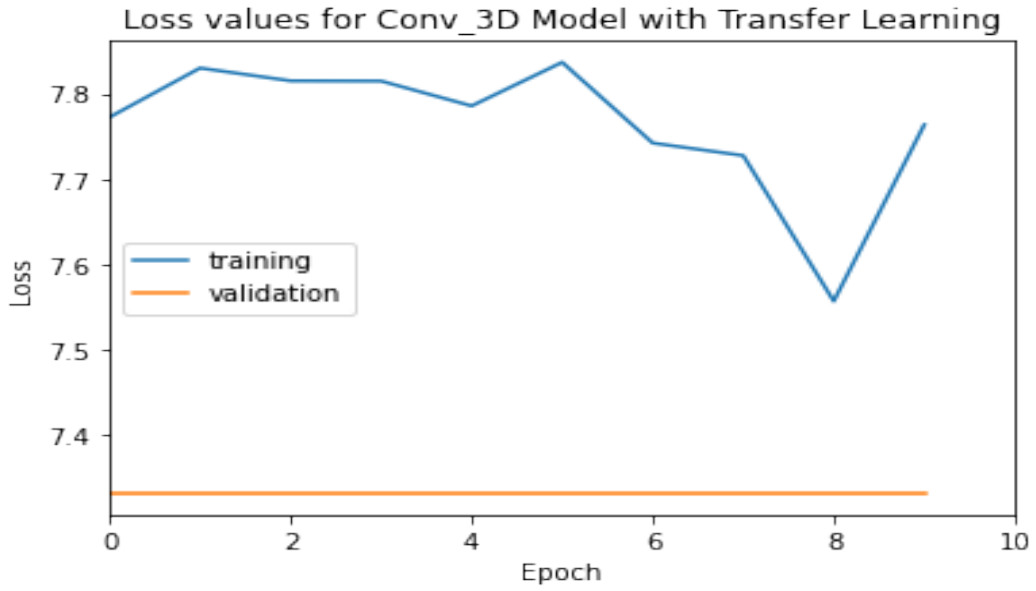

Figure 5.13: Loss Graph for Conv_3D model with transfer learning

Model	Train set	Test set
Conv_LSTM	48.5%	47.6%
Conv_3D	56%	53%
Conv_3D using Sports1M	48.6%	47.6%
Model by Shirzad et al. (2020)		70%

Table 5.2: Accuracy comparison for GIF models

5.4.4 Comparison of results for GIF models

As per the table 5.2, Conv_3D trained from scratch performs better than other models. In Shirzad et al. (2020), for GIF analysis, frames were divided and an image model was used on independent frames.

While working with independent frames, temporal information is not considered. These models take into consideration the spatial and temporal features and process them as a sequence of frames. However, the performance of the models suffers due to less number of frames per video. The models are unable to learn all the features in the video. Conv_3D model performs better as compared to other models, and a learning curve could be seen in the graphs. The Conv_LSTM model is not able to learn the features from the frames. This makes it not suitable for the purpose of GIF analysis. As GIFs are short videos, it is difficult to increase number of frames. However for training purpose, GIFs with more number of frames could be considered.

5.5 Multimodal Framework

As per results from different modals, a framework can be created for better sentiment results. The text and image modals perform better as compared to GIF modal. To calculate overall score, weighted average can be used. Higher weights

is given to image and text data and a lower or no weight can be given to GIF modal. Later, majority vote can be taken from overall results. As text and image modals have better accuracy, they will predict the sentiment better than GIF model. Moreover, the GIF modal will also contribute in the results but with lesser influence as compared to other modals.

For example, a tweet containing text, an image, and a GIF could be classified as follows:

(2 x result from text, 2 x result from image, 1 x result from GIF)

The result that occurs maximum times is considered as the final result. The majority result will be three times as compared to minority result which will be two times. This will avoid leading to ambiguous results.

The proposed method is not yet tested on a dataset. The datasets selected for this research are independent modals. The multimodal framework can be tested on Twitter data where a user has used text, image, and GIF to share their views.

Chapter 6

Conclusion and Future Work

6.1 Overall Discussion

In this research, multiple models have been implemented for Image and GIF modals. The Conv_LSTM model does not perform well on the dataset. Furthermore, it requires more computation and training time as compared to 3D CNN. Better results are achieved with the 3D CNN model. This shows that further training of the model is required with perhaps more data. Moreover, it can be seen the data is not enough to get better results. Videos with more frames can help in improving the training of the model. The analysis of spatiotemporal features indicates that there is no significant contribution in improving the accuracy of the models.

While working on the research, I have implemented distinct models for GIF sentiment prediction. In the past, there were no papers that implemented Conv_LSTM for GIF sentiment prediction. I have implemented it to check if better results could have been achieved. Conv_LSTM models encapsulate the power of convolution as well as sequence learning. Videos can be considered as a sequence of frames where LSTM can help in learning the sequence and convolution can

help in feature detection. After different experiments with the parameters of the model, the accuracy of the model is 47.6% on the test set. This suggests that the current Conv_LSTM model is not suitable for GIF analysis. However, the model could be improved by adding more layers. As layers are increased, the computation power required to train the model increases exponentially. The Transfer learning approach reduces the number of parameters to be trained and could be used with more data to get better results.

The image models have good results when trained on the data. Both the models perform well on the training and testing data. Transfer learning helps in reusing the model with original features as they are trained on millions of images. The results from 3D CNN with transfer learning approach are not good. Some changes in the layers gave better results as compared to the original architecture. The model used for transfer learning in 3D CNN was trained for action recognition. The features could be very different and thus the model could not perform well using the pretrained weights.

The VADER python library gives good results when used on the text dataset. In summary, different modals can be used in conjunction for sentiment prediction using the joint multimodal framework.

6.2 Conclusion

The overall conclusion from this research are as follows:

- A multimodal framework can be used to improve the accuracy of sentiment prediction. With the performance of models, a weighted average can help in choosing the best results from different modals.
- Performance of Conv_LSTM model suggests that it might not be suitable for the video classification task. However, adding more layers and increasing

the number of frames per video could possibly improve the results.

- Results from 3D CNN model show that results could be improved with more data. The transfer learning approach could possibly help in improving the results if more data is used in training.

- Image models give an accuracy of 87% on the training set and 79% on the test set. As compared to the VGG16 model implemented in Shirzad et al. (2020), these models have fewer parameters and get trained faster. The accuracy of models does not differ significantly as compared to VGG16 which is 83%.

6.3 Future Work

In this research, videos each with 16 frames are considered for training the model. In future work, more frames per video can be considered as it will help in better feature detection. Moreover, other feature detectors such as face recognition, and attention mechanism can be added to help concentrate on specific features in videos. Similarly, attention can also be implemented on image models to detect specific features in the images. Targeting features in the model could give better overall results. The multimodal framework can be tested on the Twitter data where users have used text, images, and GIFs in the posts.

References

Saeideh Bakhshi, David A. Shamma, Lyndon Kennedy, Yale Song, Paloma de Juan, and Joseph 'Jofish' Kaye. Fast, cheap, and good: Why animated gifs engage us. CHI '16, page 575–586, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858532. URL https://doi.org/10.1145/2858036.2858532. 1

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 223–232, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324045. doi: 10.1145/2502081.2502282. URL https://doi.org/10.1145/2502081.2502282. 7

Weixuan Chen and Rosalind Picard. Predicting perceived emotions in animated gifs with 3d convolutional neural networks. pages 367–368, 12 2016. doi: 10.1109/ISM.2016.0081. 10, 11

Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 20

Michael Gygli and Mohammad Soleymani. Analyzing and predicting gif interestingness. In *Proceedings of the 24th ACM International Conference on Mul-*

timedia, MM '16, page 122–126, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450336031. doi: 10.1145/2964284.2967195. URL https://doi.org/10.1145/2964284.2967195. 1

C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550. 22

Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. Predicting viewer perceived emotions in animated gifs. MM '14, page 213–216, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330633. doi: 10.1145/2647868.2656408. URL https://doi.org/10.1145/2647868.2656408. 6

Taeyong Kim and Bowon Lee. Multi-attention multimodal sentiment analysis. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 436–441, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370875. doi: 10.1145/3372278.3390698. URL https://doi.org/10.1145/3372278.3390698. 9

Chao Li, Shouqian Sun, Xin Min, Wenqian Lin, Binling Nie, and Xianfu Zhang. End-to-end learning of deep convolutional neural network for 3d human action recognition. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 609–612, 2017. doi: 10.1109/ICMEW.2017.8026281. 19

Tianliang Liu, Junwei Wan, Xiubin Dai, Feng Liu, Quanzeng You, and Jiebo Luo. Sentiment recognition for short annotated gifs using visual-textual fusion. *IEEE*

Transactions on Multimedia, 22(4):1098–1110, 2020. doi: 10.1109/TMM.2019.2936805. 1

Lokesh Mandloi and Ruchi Patel. Twitter sentiments analysis using machine learninig methods. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–5, 2020. doi: 10.1109/INCET49848.2020.9154183. 5

Shervin Minaee and AmirAli Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *CoRR*, abs/1902.01019, 2019. URL `http://arxiv.org/abs/1902.01019`. 9

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2015. doi: 10.1109/CVPR.2015.7298687. 14

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. 21

Amirhossein Shirzad, Hadi Zare, and Mehdi Teimouri. Deep learning approach for text, image, and gif multimodal sentiment analysis. In *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 419–424, 2020. doi: 10.1109/ICCKE50421.2020.9303676. 8, 29, 38, 42

Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9:p9420, 10 2019. doi: 10.29322/IJSRP.9.10.2019.p9420. 5, 8

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri.

Learning spatiotemporal features with 3d convolutional networks, 2015. 19, 32, 36

Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317, Oct 2017. doi: 10.1109/ICCVW.2017.45. 16

SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 18

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks, 2015. 10, 11

Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: Image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323321. doi: 10.1145/2502069.2502079. URL `https://doi.org/10.1145/2502069.2502079`. 1

Hafed Zarzour, Bashar Al shboul, Mahmoud Al-Ayyoub, and Yaser Jararweh. Sentiment analysis based on deep learning methods for explainable recommendations with reviews. In *2021 12th International Conference on Information and Communication Systems (ICICS)*, pages 452–456, 2021. doi: 10.1109/ICICS52457.2021.9464601. 5, 6, 11

Appendix A

Code

The code written as part of this research is available on Github on the below link.

`https://github.com/tjbohari/multimodal-sentiment-analysis`