A Novel Image Manipulation - Polarity based approach to Fake Content Detection in OSN's

A Project Report

submitted by

PARTH MAHESHKUMAR PATEL, HARMANDEEP SINGH

(COE17B020, COE17B024)

in partial fulfilment of requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

Department of Computer Science and Engineering

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING KANCHEEPURAM

MAY 2021

DECLARATION OF ORIGINALITY

We, **Parth Maheshkumar Patel, Harmandeep Singh**, with Roll No: **COE17B020, COE17B024** hereby declare that the material presented in the Project Report titled **A Novel Image Manipulation - Polarity based approach to Fake Content Detection in OSN's** represents original work carried out by us in the **Department of Computer Science and Engineering** at the Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram.

With our signature, We certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Parth Maheshkumar Patel, Harmandeep Singh

Place: Chennai

Date: 06.05.2021

CERTIFICATE

This is to certify that the report titled **A Novel Image Manipulation - Polarity based approach to Fake Content Detection in OSN's**, submitted by **Parth Maheshkumar Patel, Harmandeep Singh (COE17B020, COE17B024)**, to the Indian Institute of Information Technology, Design and Manufacturing Kancheepuram, for the award of the degree of **BACHELOR OF TECHNOLOGY** is a bona fide record of the work done by him/her under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr B Sivaselvan

Digitally signed by Dr B Sivaselvan
DN: cn=Dr B Sivaselvan, o=IIITDM Kancheepuram, ou=Dept of CSE, email=sivaselvanb@iiitdm.ac.in, c=IN
Date: 2021.05.04 19:52:53 +05'30'

Dr. B Sivaselvan
Project Guide
Assistant Professor
Department of Computer Science and Engineering
IIITDM Kancheepuram, 600 127

Place: Chennai

Date: 06.05.2021

ACKNOWLEDGEMENTS

We would extend our sincerest gratitude to our advisor **Dr B Sivaselvan** for helping in choosing an interesting area to work on and for guiding us.

We express our sincere thanks to PhD Scholar **Mr. Santosh Kumar Uppada**, for giving necessary advice and guidance through the period.

We would also like to thank **Ms. Harini R (Roll No: COE16B018)** whose research on Credibility of Visual Information on Online Social Networks were a big help for this project and also providing the relevant research papers.

ABSTRACT

Online Social Networks (OSNs) have billions of active users today. With huge volumes of information is being shared online, the credibility of such information falls into question - one scenario being the widespread evil of ‘fake’ news. Fake news, with their rapid spread, has had, and continues to have disastrous effects on society today, creating fear, panic, violent clashes among the public, and also leading to events which have major global impact (Billion dollar stock value wipeouts, US Presidential elections in 2016, Brexit, to name a few). Fake news is easy to produce and post on online social networks, making it difficult to control its spread. The "believability" of the news material, the number and "influential influence" of the user accounts that post such content, psychological aspects such as confirmation bias, backfire impact, and so on are all factors that contribute to the dissemination of fake news. With the advancement of multimedia technology, more and more social media news now includes content in a variety of formats, such as text, images, and videos. The manipulation of photos, text, and videos is common in fake news posts, suggesting the need for a multimodal fake news detection system. Visual information such as images in fake news posts, termed as ‘fake images’, have been found to significantly increase the ‘believability’ of the post and thereby fuel its spread, often misleading naive users. Such fake images could either be tampered or it can be real images which are inappropriately used with malicious intent. Identifying such fake images in the immense inflow of content posted online becomes a challenging task.

In this project, we propose a **multi-modal** framework to flag fake posts efficiently among a wider pool of both fake and genuine posts on online social networks. It exploits features unique to both the images and text in fake news posts. The proposed model consists of two CNN architectures to learn visual features and one language model (BERT) to learn text features. **The model utilizes the spatial properties of CNNs to look for physical alterations in an image as well as analyse if the image reflects a negative sentiment, since fake images often exhibit either one or both character-**

istics. Standard fake news datasets and image polarity datasets collected from popular social networking sites such as Reddit and Flickr are used for training. Thus, unlike traditional image forensic techniques, **the model has been trained to identify both tampered and untampered but misleading images. The proposed model performs better than the baseline, having an accuracy of 91.94% with an Precison, Recall and F1 score of 93%.**

KEYWORDS: Fake News Detection; Multimedia; Online Social Networks; Natural Language Processing; Deep Learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CNN	**C**onvolutional **N**eural **N**etwork
NLP	**N**atural **L**anguage **P**rocessing
ANP	**A**djective **N**oun **P**air
RGB	**R**ed **G**reen **B**lue
SVM	**S**upport **V**ector **M**achine
AMT	**A**mazon **M**echanical **T**urk
LSTM	**L**ong **S**hort **T**erm **M**emory
GRU	**G**ated **R**ecurrent **U**nit
BiLSTM	**Bi**directional **L**ong **S**hort **T**erm **M**emory
BiGRU	**Bi**directional **G**ated **R**ecurrent **U**nit
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers

CHAPTER 1

Introduction

Social Networks, such as Facebook, Instagram, and Twitter, are a prevalent source of information in today's world. Facebook has 2.45 billion monthly active users as of the third quarter of 2019. Users share information on such platforms and interact with others. With an enormous amount of information being shared and spread rapidly on such networks, the credibility of such content is an important aspect to evaluate and it becomes the ample target of people who wish to spread misinformation and propaganda.

1.1 Fake News

The concept of "fake news" is not recent. It had been present in society for a long time, but the harm it caused mankind made it a significant problem that needed to be tackled by the scientific community. The rapid dissemination of content on Online Social Networks, while carrying multiple benefits - enables users to be aware of current happenings and share their opinions, instant messaging and entertainment, et cetera - has also paved way to widespread evil of 'Fake news'. Shu et. al define fake news as a news article that is intentionally and verifiably false. [8]. These are online posts with false information which misleads users and instigates fear and chaos. Such posts are intentionally written to mislead users into believing that they are true, making their detection challenging - and manual fact-checking is extremely time taking, making it an ineffective method to verify the enormous volume of posts published every day ($\sim$500 million tweets per day on Twitter).

1.2 Spread of 'Fake News'

The rapid creation and widespread dissemination of fake news on online social media sites can be attributed to a number of factors. Firstly, it is cheaper and quicker to publish

fake news posts on social media rather than on traditional mediums such as television and newspapers. There is also the 'power law' observed in such social media that posts can spread more quickly and reach wider audiences if the posts target a few influential people in the online social network.

Psychological factors play a significant role in the dissemination of false news as well. Because of the way news appears on their feed/homepage, users often interact with only certain types of news. Users also tend to form groups with other like-minded people, polarizing their opinions. [9] This *echo chamber effect* observed in social networks shows that belief and biased information is often amplified. [10] There is also *confirmation bias* - people tend to trust fake news if it aligns with their pre-existing knowledge.[11] Users who engage with fake news posts can be *malicious* users who spread the false information intentionally and naive users who participate unintentionally, driven by influence and psychological factors. [12]

1.3 Motivation and Scope

1.3.1 Need for a Fake News Detection System

Social networking sites have grown in importance as a source of information and a medium for people to freely express themselves. However, the convenience and openness of social media facilitated the spread of fake news, or news that contains deliberately misleading facts, which not only undermined cyberspace order but also harmed real events around the world. [13]

For example, during the final three months of the 2016 presidential election in the United States, fake news created to favour one of the two candidates was commonly believed and shared over 37 million times on Facebook and Twitter.(refer Fig. 1.1) [8].

In the social realm, hundreds of innocent people were beaten to death by locals in India as a result of widely disseminated false news about child trafficking on social media. In the current scenario (as of May 2021) of the COVID-19 pandemic, with rapidly changing living conditions and governmental policies, the rise of misinformation is un-

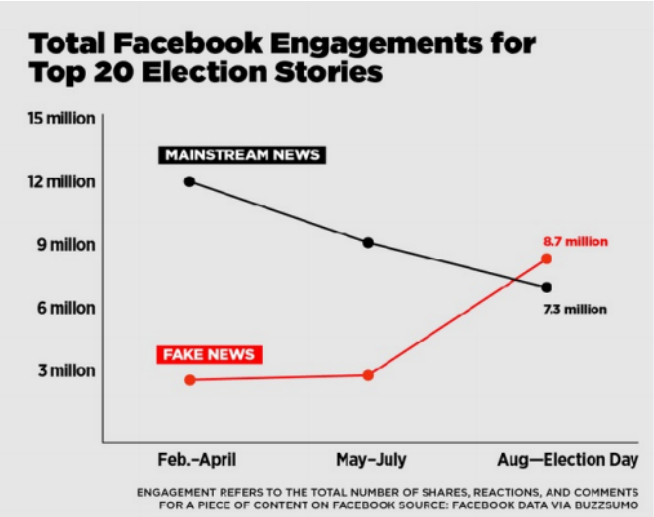


Figure 1.1: Buzz feed analysis shows how Viral Fake Election News Stories outperformed Real News on Facebook [1]

precedented. Fact-checking website Boom reports that the number of false/misleading claims and misinformation has a positive correlation with the number of COVID-19 cases in the country. It has also been found that about 6**5% of COVID-19 related misinformation were shared with multimedia such as images and videos.** A majority of COVID-19 news has been found to originate from Facebook, WhatsApp and Twitter, the major social networking platforms in the country.[14]

Figure 1.2: A particular instance of fake news in which it is claimed that cocaine consumption can cure the 2019 corona virus outbreak in China. [2]

The widespread dissemination of fake news can have serious negative consequences for individuals and society, such as:

- Can break the news ecosystem's authenticity balance.
- It is designed to persuade customers to accept prejudiced or false beliefs.
- Affects how people interpret and react to actual news.

This makes detecting fake news a crucial task. An automated fake news detection system is needed for reducing the widespread evil of 'fake' news. [15]

1.3.2 Need for a Multimodal Fake News Detection System

With the development of multimedia technology, more and more social media news contains information with different modalities, e.g., texts, pictures and videos. Photos and videos are more appealing and popular than plain text, resulting in increased news dissemination. Tweets with videos, for example, posts with images receive 18 percent more clicks, 89 percent more likes, and 150 percent more retweets than those without. Unfortunately, fake news also exploits this advantage. In order to attract and confuse readers, fake news often includes misrepresented or even tampered images or videos. [13]. The manipulation of photos, text, and videos is common in fake news posts, suggesting the need for a multimodal fake news detection system.

Given that fake news images online significantly contribute to the propagation of false information, this project was devised to analyse such images and identify features to distinguish images used in fake news posts from those in genuine ones. Fake news images can be manipulated versions of real ones or can even be unedited images which are misappropriated. Thus, traditional forensic techniques are inadequate to handle the diverse nature of fake images on social media. There is a need to develop a framework which can effectively learn useful features from the varied nature of images in fake news to be able to distinguish them from those in real posts. Such a framework could hugely benefit online social networks in their efforts to curb the proliferation of fake news on their platforms.

1.4 Characteristics of Fake images

'Fake' images can be of two kinds. **Tampered images** are those which have been digitally modified to manipulate viewers. **Misleading images** are those which are actually real, unaltered images, but are embedded in inappropriate contexts.These include images of an earlier event being shared as a current scenario, or even images which are

misrepresented with corrupt intent.[16, 17]
It has been observed that *fake images* are often eye-catching and carry emotional impact. Thus, it becomes necessary to map such psychological triggers to characteristics of the image. These psychological patterns are not just limited to direct visual appearance and are beyond the common object-level features. Hence traditional image sets are not suitable for this task of fake image classification. Gathering large labelled datasets containing posts with real and fake images is difficult because human verification and labelling of posts is time-taking and not fast enough to deal with the immense data online. In subsequent sections, this report discusses notable strides in this endeavour of identifying fake images as manipulated and about the image emotion.

1.5 Image Manipulation Detection

The use of image editing tools, such as Adobe Photoshop, GNU image manipulation programmes (GIMP), Affinity Photo, Paintshop, etc., has now made manipulation of image very convenient. A large number of images are generated, and of the ease with which computer software or mobile apps are available, anyone can easily tamper with them. Images are tampered with without leaving any visible traces thanks to the rapid development of image editing tools.Humans have a limited capacity to differentiate between the original and tampered image, according to one report.
These manipulated photos may be used as evidence in a criminal case, to defame a person's character, or, more recently, to spread false news and rumours.[3]

1.5.1 Forensic Methods

Lago et. al comment that image forensics methods are generally not be robust enough to verify fake images, because images uploaded and shared on online social networks are compressed and resized, often multiple times. This erases the traces which image forensics methods look for. [18]

Metadata of an image, such as date and time of when it was taken, location depicted, etc, can be useful in verifying the *truthfulness* of the image. However, such analysis is

Figure 1.3: A well-known image manipulation example, the composite photo of Senator Millard Tyding and American Communist Party Leader Earl Browder (left)[3]

likely to be too slow to perform given the rapid spread of news online. Image manipulations including splicing (moving an object from one image to another) and copy-move (moving an object from the same image to a different location) leave digital traces that forensics methods attempt to detect.

Some techniques to detect tampering of images are:

- Error Level Analysis
- JPEG Ghosts
- Block Artifact Grid Detection
- Median-filter noise residue analysis
- Double Quantization Likelihood Map
- Color Filter Array Interpolation Pattern
- Analysis of quantization tables, thumbnails
- Photo Response Non-Uniformity

1.5.2 Advanced Methods

- Deep Neural Networks can be used to detect image manipulation, without the need for a feature extraction task; feature extraction and classification can be done in an end-to-end process. Convolution Neural Networks (CNNs) have been

shown to be extremely efficient in detecting image manipulation. The drawback, however, require large labelled dataset for training. [19, 20, 21]

- *Splicebuster* focuses on splice detection, which is performed on a single image without reference/apriori information. In short, *Splicebuster* works by extracting local features related to the co-occurrences of quantized high-passed residuals of the image. [22]

However, with various kinds of image manipulations found in fake images on social media and also with the case of misleading images, this method would not prove effective for the task of fake image classification.

The next chapter discusses papers that leverage neural networks to achieve fake news identification.

1.6 Image Polarity Detection

Image Polarity detection deals with understanding how an image affects people - whether it invokes positive emotions or negative sentiments. Similar to sentiment classification of an input text (e.g., taken from a review, a comment or a social post) in terms of positive, negative or neutral polarity.

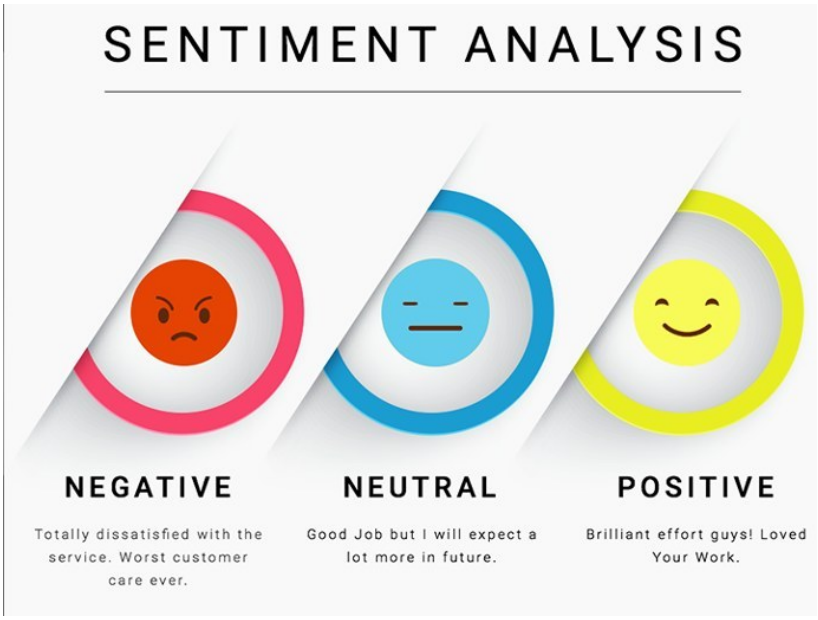


Figure 1.4: Sentiment analysis on text

Text sentiment analysis techniques have been studied in the past in various contexts

(e.g., social network posts analysis, product reviews, political preferences, etc.). Sentiment analysis for visual content is a relatively new field, with a variety of techniques emerging in recent years.

This finds strong relevance to the context of online social networks because, apart from short texts, people tend to use images to express their emotion and experiences on these online platforms. The visual content not only carries semantic information such as object or scene portrayed but also accompanies the sentiments and emotions. Thus, sharing images and videos on the online social network has become a popular medium of expression. [23]

This domain was identified as a potential direction for this project during the course of articulating what contributes to the 'fakeness' of news images in social media. Fake news images often carry strong visual impact and invoke negative emotions in the viewer. An image could be digitally altered to bring about such emotions, or even unaltered images could be used in a misleading context as discussed above. Thus, analysing the sentiment of the image helps understand the emotional impact carried by the image could help identify whether it is a fake news image or not.

Key research works in this area have been discussed in chapter 2.

CHAPTER 2

Literature Survey

Given a post with its accompanying image on an online social network, methods discussed by ongoing research to assess whether the image conveys false information, ie, whether the image is real or fake are discussed in this section.

2.1 Introduction

The problem of fake news/posts is multifaceted - involving text and multimedia contained in the post, the user account which uploaded the post, the number of engagements(shares,comments) the post acquires, and also psychological aspects of rumor spread.[24] Most literature in this area focus on text and social context. [25, 26, 27, 28] Some works focus on utilising the visual modality, with some using forensic techniques [6, 18] and some using pre-trained networks [29, 30, 31]. Qi et. al propose a multi-branch CNN-RNN to extract semantic information from images.[17]

For the purposes of this report, the latest key research papers closely related to the scope of this project shall be discussed below and other papers relevant to the topic/method discussed shall be cited accordingly.

2.2 A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer (2016)

Bayar et. al propose a modified convolutional neural network architecture to learn image manipulation features. This aims to address the inadequacy of using traditional

forensic techniques - an image can be manipulated in multiple ways, using combinations of tampering methods, thus a traditional forensic examination would require testing the image for each of the methods. Convolutional Neural Networks learn features about an image's content rather than those corresponding to image manipulation detection. Therefore, the paper proposes a new form of a convolutional layer that suppresses an image's content and rather learn image manipulation detection features. They state that image manipulations alter the local structural relationships between pixels, and thus the modified CNN architecture aims to learn the relationship between a pixel and its local neighbourhood. The proposed architecture contains 8 layers with ReLu activations and can automatically detect different manipulations with an average accuracy of 99.10[32] As discussed above, apart from manipulated images, fake news images can also be unedited, untampered images which are used for misleading purposes. Therefore the proposed method aims to learn features relevant to both types of fake images found on online social networks.

2.3 Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs (2017)

Jin et. al propose a Recurrent Neural Network with an attention mechanism (att-RNN) to fuse multimodal features for rumor detection. [30] LSTM is used to obtain a joint representation of text and social context. The test is represented as a word embedding vector, where the embedding vector for each word is obtained with a deep network. Features of social contexts are inferred from text semantic features, hashtags, and user engagement parameters such as mentions and retweets. **For visual features, the output of the second to last layer of the VGG19 architecture pretrained on ImageNet is used.** RNN is used to fuse the text and social context. The attention mechanism adjusts the visual representation according to the RNN output, with the aim of assigning more weights to visual neurons which have similar semantic meaning with the text.

The focus of this paper is on the fusion of different modalities of a post rather than extracting features of fakeness/genuity of images in them. The accuracy achieved by att-RNN based on the visual modality alone on the Weibo dataset is 60.8

2.4 EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection (2018)

Wang et. al propose EANN to derive event-invariant features which consists of a multimodal feature extractor, fake news detector and an event discriminator. [31] For fake news detection, the feature extractor and fake news detector work together to learn discriminable representation. The event discriminator removes event-specific features and retains only shared features across events. This is done to address the issue of identifying fake news on newly emerged events.

The fake news detector acts on text and visual features and identifies if the post is fake or not. The event discriminator measures the dissimilarities of the feature of different events and capture event invariant feature representations. Textual features are extracted via a modified CNN. Similar to att-RNN, E**ANN uses a pre-trained VGG19 to obtain the visual features** (Thereby no available baseline from EANN for the visual modality alone).

However, in this project, the proposed aims to extract visual features relevant to the task of fake image identification .

2.5 MVAE: Multimodal Variational Autoencoder for Fake News Detection (2019)

Dhruv et. al propose MVAE to learn a shared representation of multimodal information for fake news detection. MVAE employs a binary classifier and a bimodal variational autoencoder. [29] MVAE has an encoder to encode information from text and image into a vector, decoder to reconstruct back the original image and text from the vector and a fake news detector module which uses the representation to predict if the given news is fake or not. The textual encoder uses RNNs with LSTM cells which help store information over long time periods. For visual information, here too, the **outputs of a pretrained VGG19 architecture trained over ImageNet are used.** (Thereby no available baseline from MVAE for the visual modality alone) The decoder reconstructs

data from the multimodal representation, and the fake news detector classifies the post as fake news or not.

2.6 Exploiting Multi-Domain Visual Information for Fake News Detection (2019)

Qi. et al propose Multi-domain Visual Neural Network(MVNN) to fuse visual information of frequency and pixel domains for fake news detection. [17] It is stated that fake images, after being uploaded and downloaded multiple times, the images have heavy re-compression artifacts such as block effect. Re-compressed and tampered images present periodicity in the frequency domain, while from the pixel domain, the visual impacts and emotional provocations often found in fake news are observed. For capturing information from the frequency domain, the DCT coefficients of the image are sent as input to a CNN network. For the pixel domain, MVNN contains a CNN-RNN network to extract features of different semantic levels. This is a multi-branch CNN which utilizes a bidirectional GRU(Bi-GRU) network. Both physical and semantic features of images are fused via the fusion sub-network. Since not all features contribute equally to the fake images detection task, an attention mechanism is employed. Qi et. al report an accuracy of **84.6%** via this approach.

2.7 Analyzing and predicting sentiment of images on the social web (2010)

The authors of this paper study the relationship between image sentiment and visual content. In order to derive the sentiment of the images, the metadata of the image (e.g., image title, description, tags) is used and the lexical resource SentiWordNet[33] is used as a dictionary of positive and negative words. A positive sentiment label is assigned to image if the metadata majorly contains words which carry a positive sentiment and likewise is done for labelling negative sentiment images. Visual features are obtained from global RGB histograms and visual bag of words model. The bag of words model

enables an image to be described with terms. The terms are treated as completely independent of each other, with no bearing with their relative or absolute positions in the image. Each bin of the RGB histogram can be considered to represent a particular term, which in turn represents a range of similar colours. In general, the global color histogram does not carry any information about the layouts of the colours in the image. Therefore, the authors split each image into blocks and compute the RGB histogram for each block. This work showed that there are strong dependencies between the sentiment scores of metadata of the image and the visual features extracted from the image. [34]

2.8 DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks (2014)

Chen et. al introduce a deep convolutional neural network based method for visual sentiment analysis. Adjective Noun pairs (ANP) discovered from the tags of the images are used as sentiment labels for the images. In this work, a CNN model pre-trained on ImageNet dataset is fine-tuned for the purpose of sentiment analysis. This has been found to be better than previous approaches which used SVM classification methods.[35]

2.9 Robust image sentiment analysis using progressively trained and domain transferred deep networks (2015)

In this paper, a progressive approach is used to train a CNN (Progressive CNN or PCNN) to conduct visual sentiment analysis and mark images as 'positive' or 'negative.' A subset of training images with high prediction scores is chosen by the algorithm. This is done to address the weakly labelled nature of the dataset and the authors suggest that is effective to improve the generalizability of the model. [36]

The authors also a build a new dataset from images obtained from tweets by employing crowd intelligence on the Amazon Mechanical Turk (AMT) platform. 5 AMT

workers are asked to label about 1000 images as that evoking positive or negative sentiments, and those images are retained where all 5 AMT workers agree on the same label.

2.10 Discovering affective regions in deep convolutional neural networks for visual sentiment prediction (2016)

In this work, the authors start with the fact that both global distributions and salient objects carry massive sentiments. Sun et. al proposed a deep framework algorithm for discovering affective regions (ARs). An AR is a "salient" region that contains one or more objects, which could attract the user's attention mostly. Also the ARs carry massive emotion. An off-the-shelf tool was used to generate N object proposals from a query image and they were ranked based on their objectness scores. The sentiment score for each proposal is then calculated using a pre-trained and fine-tuned CNN model.

The authors then combined both scores and choose the top K regions from a pool of N candidates. These K regions are thought to be the most affective in the input image. Finally, deep features are extracted from the entire image and selected regions, and sentiment labels are predicted. The results of the experiments show that this method can detect affective local regions and achieving state-of-the-art performances on several popular datasets.

2.11 Multimodal sentiment analysis: A multitask learning approach (2019)

Fortin et. al ppresent a multitask approach for sentiment analysis that combines multimodal content such as images and text. The model comprises a text classifier, an image classifier, and a classifier that predicts by combining the two modalities. The framework developed is able to handle cases where either one modality is missing. Experiments have shown that using a multi-task learning method act as a regularisation tool that can

help with generalisation. For the visual network, Densenet-121 pretrained on ImageNet is used. For the textual network, the input text is embedded using pre-trained representations of words in 300 dimensions. To predict the class, the multimodal classifier uses two completely connected layers and a softmax layer.

2.12 Inferences

- Forensic techniques alone are insufficient to handle the problem of fake news image identification, and a universal approach needs to be developed which can handle the scale and varied nature of fake news images.
- Most works focus on fusing multi-modal information such as text, social context, and images. The output of a pretrained VGG-19 architecture trained on ImageNet is commonly used among these works. However, this lacks information unique to fake news images detection, which has varied characteristics (physical, semantics as mentioned above). Therefore, there is a need to develop a framework which focuses on identifying features unique to fake news images to aid in their identification.
- In the area of image polarity detection, we see a shift towards CNN methodologies to obtain sentiment cues from the image.
- The early methods were mostly based on hand crafted visual features, more recent approaches exploit textual metadata and features learned directly from the raw image (i.e., CNN based representations).
- There is also significant amount of work in using multiple modalities for sentiment analysis. However, different research works use varied datasets and hence comparison of performances of various frameworks developed is challenging.
- Previous research has shown that combining features from different layers can help achieve better output when using CNN to learn high-level semantic representations, particularly in tasks like salient object detection and image emotion classification.

CHAPTER 3

Problem Statement

In the previous chapters, the problem area was analysed and related literature was presented. Here, the refined problem statement and the various challenges identified shall be discussed.

3.1 Fake News Detection based on Image caption

As previously stated, false news is used to manipulate the audience, and for this, they use a specific language to attract readers. Non-fake news, on the other hand, would usually be moved to a different language list, as it is more formal. As a result, it is fair to identify fake news using linguistic features that capture various writing styles and sensational headlines.

3.2 Fake News Detection based on Image Manipulation

Reviewing the problem of fake news online and its widespread negative effects, it is found that the visual contents of such fake news significantly contributes to the spread of the news. This has been attributed to the observation that images offer a perception of 'reality' and hence users are often easily misled by fake news with visual content embedded.

Traditional forensic methods and most work on fake news which focus on fusing multiple modalities miss to model and capture visual characteristics unique to fake news images found online.

3.3 Fake News Detection based on Image Sentiment

This domain was identified as a potential direction for this project during the course of articulating what contributes to the 'fakeness' of news images in social media. Fake images have been attributed to the emotional impact and negative sentiments the image evokes in the viewer. By offering a perception of reality, the fake news image influences the viewer to share the piece of news - leading to the fake news being spread across the online social network.
Sometimes an older or unrelated image, which is unaltered is used to spread the propaganda and thus it is important to find out what type of an emotion does that image carries, i.e. religious , unbiased , person centric, sad, violent etc.

This project aims to focus on both the textual and visual modality and model features unique to fake news, as well as to identify posts that have been manipulated and are not factually true, in order to aid in the identification of fake news posts on online social networks.

Thus the problem statement: **Given an online post containing both text and image, identify if the post is 'fake' or real (genuine), i.e., predict whether the post has been used for fake news purposes or not. A fake post can be tampered/manipulated version of genuine post or can be untampered but misleading.**

3.4 Challenges

The initial challenges identified after doing the literature survey addressing the problem statement are as follows:

- Obtaining a large dataset which contain the ground truth labels of posts (fake or real) is difficult, since manual fact-checking is extremely time consuming.
- Fake news can vary from the reality in terms of writing style and quality (according to the Undeutsch hypothesis), quantity (according to the information manipulation theory), and expressed sentiments (according to four-factor theory).[12] The framework developed needs to capture those features for fake news text which can help distinguish fake text from genuine ones.
- Images uploaded online are often of varying sizes and quality. They originate

from different camera models. Thus, the framework developed needs to be robust to these variations and capture features invariant of size, quality or camera model.

- Images online (both real and fake) often have text attached to them. Some images are also collages of other images. This makes the information noisy and hinders the process of capturing visual features.
- Fake images/text are of two kinds - tampered and misleading. The framework developed needs to capture those features for both variants for fake images/texts, which can help distinguish fake images/texts from genuine ones. Identifying such features is challenging because fake images are deliberately made very 'believable', i.e., they pretend to be genuine.
- For tampered fake images/texts, the original genuine version is unavailable, hence we do not have a reference image/text to compare with the potentially fake image/text and identify alterations and manipulations.
- Even real images may contain physical alterations which are not malicious - such as changing brightness or contrast, resizing and re-scaling the image, cropping, rotations, applying filters to sharpen or blur the image, adding colour hues, etc.
- There are new events and trends emerging on a daily basis on online social networks. Images used in these platforms depict and convey information about specific events. Thereby their visual features tend to be closely tied with the event. Therefore the analysis on such past history of images should aid the framework be adept in handling new event images. In other words, the framework needs to be generalisable to new emerging events never seen earlier.

In subsequent chapters, the methodology and approach used by this project will be discussed, followed by in-depth discussions of the project's progress and bottlenecks at each stage.

CHAPTER 4

Dataset used for Fake News Detection

Manually assessing the veracity of news normally necessitates domain experts who conduct in-depth review of claims, additional evidence, context, and reports from reliable sources.

Given that manual labelling is challenging, we make use of the publicly available dataset to evaluate our architecture for fake news detection. Our model training is carried out on Fakeddit dataset, which is publicly available.

4.1 Fakeddit dataset

Fakeddit is a large-scale multimodal fake news dataset with over 1 million samples containing text, image, metadata, and comments data gathered from a wide range of sources.

The fake news articles in this dataset are scraped from Reddit, a social news and discussion platform where users can submit submissions to various subreddits. Reddit is one of the top 20 most visited websites on the internet. Over 1 million submissions from 22 separate subreddits make up Fakeddit. The earliest submission was on March 19, 2008, and the most recent submission was on October 24, 2019. Submissions were collected using the pushshift.io API8. The authors collected the submission title and image, as well as comments made by users who interacted with the submission, as well as other submission metadata such as the score, the author's username, the subreddit source, the sourced domain, the number of comments, and the up-vote to down-vote ratio. Multiple quality assurance steps are used to fine-tune the samples.

Users may use the dataset for 2-way, 3-way, and 6-way classification since each data sample contains several labels. This allows for both high-level and fine-grained

classification of false news. The 2-way classification is used to decide whether a sample is real or not. The three-way classification decides whether a sample is absolutely real, is fake but contains true text (e.g., direct quotes from propaganda posters), or is fake with false text. The 6-way classification decides whether a sample is entirely real, satirical/parody, or deceptive, imposter, and manipulated material, or a false relation. In Figure 4.1, we show examples from each class for 6-way classification.

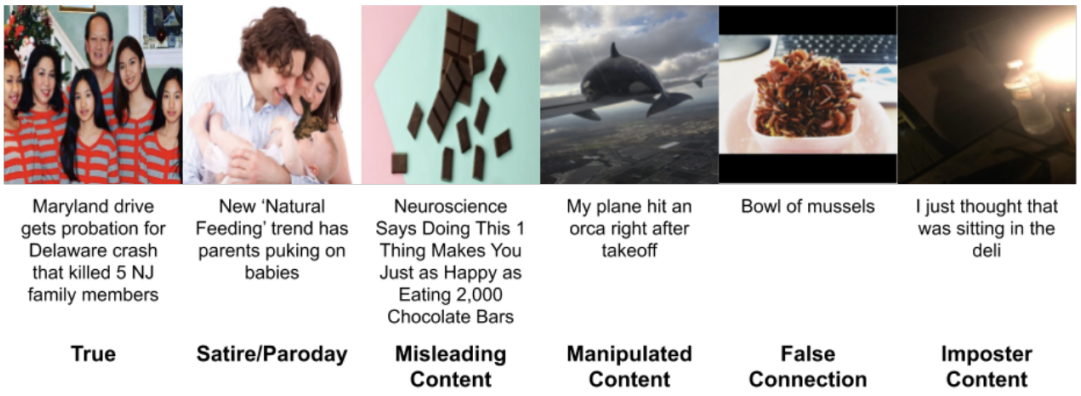


Figure 4.1: Examples of dataset with 6-way classification labels.

In our dataset, approximately 64% of the samples contain both text and images. Our studies are based on these multimodal samples. The dataset is split into training, validation and test sets. The statistics of the dataset are as follows:

	Training	Validation	Test
Real samples	222081	23320	23507
Fake Samples	341919	36022	35812
Total	564000	59342	59319

Table 4.1: Statistics of Fakeddit dataset

Examples of samples in dataset

Figure 4.2 shows two fake images and Table 4.2 show their corresponding text and metadata from the dataset and Figure 4.3 and Table 4.3 shows that from the real images category

	Sample 1	Sample 2
author	hakiku	SpaceFloow
clean_title	hawaiian terrorists in after brutally murdering a couple because they threw pineapple pizza in the garbage	obamas plan to fix the economy
created_utc	1.5E+09	1.38E+09
domain	i.redd.it	-
hasImage	TRUE	TRUE
id	6mxyac	cctlmvk
image_url	https://preview.redd.it/m112k0e6299z.jpg?width=320&crop=smart&auto=webp&s=83612b2820117a966fc3f391352425680a0036e6	http://i.imgur.com/ipqgr0c.jpg
linked_submission_id	-	1oluv5
num_comments	1	-
score	8	63
subreddit	fakehistory	psbattle_artwork
title	Hawaiian terrorists in after brutally murdering a couple because they threw pineapple pizza in the garbage (1995)	Obama's plan to fix the economy
upvote_ratio	0.8	-
2_way_label	0	0
3_way_label	2	2
6_way_label	2	4

Table 4.2: Text and metadata of fake news.

	Sample 1	Sample 2
author	likeliqor	RanchDogTheBand
clean_title	this cap i saw in bangkok has a solarpowered fan on its brim	the moon looks like just another antenna
created_utc	1.57E+09	1.56E+09
domain	i.redd.it	i.imgur.com
hasImage	TRUE	TRUE
id	cr603j	c1qcm9
image_url	https://preview.redd.it/dxnaodbg6tg31.jpg?width=320&crop=smart&auto=webp&s=eea4cd797b7df09cd86c3d292bb920716d0a9ce5	https://external-preview.redd.it/4Cw84J9rpVplYc7YerINFTJB3cP_mgLCqlx9k9ABihs.jpg?=320&crop=smart&auto=webp&s=1d7ba84478162871ed2b9c4810ea37d370324215
linked_submission_id	-	-
num_comments	9	2
score	43	9
subreddit	mildlyinteresting	mildlyinteresting
title	This cap I saw in Bangkok has a solar-powered fan on its brim	The moon looks like just another antenna
upvote_ratio	0.86	0.74
2_way_label	1	1
3_way_label	0	0
6_way_label	0	0

Table 4.3: Text and metadata of real news.

Figure 4.2: Examples of fake images in training set

Figure 4.3: Examples of real images in training set

The following chapters describes how we used this dataset in various unimodal and multimodal approaches for detecting fake news.

CHAPTER 5

Fake News Detection based on Image caption

Fake news potentially differs from the truth in terms of writing style and quality, quantity such as word counts, and sentiments expressed. We implemented and evaluated various text modality models. The models were fine-tuned on the Fakeddit dataset and used only the textual information (image caption) in posts to determine whether they were fake or not

5.1 LSTM + CNN

The first layer is the Embedded layer, which represents each word with 100 length vectors. (similar to Fig 5.1.) I used word embeddings from pre-trained Glove.

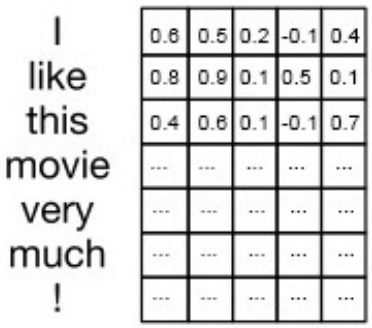


Figure 5.1: Word Embedding matrix

Convolutional neural networks are particularly good at detecting spatial structure in data. An LSTM layer can then learn sequences from the learned spatial features. After the Embedding layer, we can easily add a one-dimensional CNN and max pooling layers, which will feed the merged features to the LSTM. The following layer is the LSTM layer, which has 100 memory units. There are various Dense layers along with Dropout and batch normalization layers. Finally, we use a Dense output layer with a 2 neuron and a softmax activation function.

Trained the model for 20 epochs with batch size 1024, Adam optimizer, binary crossentropy loss and 3 callback functions – [CSV logger, tensorboard, Model Checkpoint]

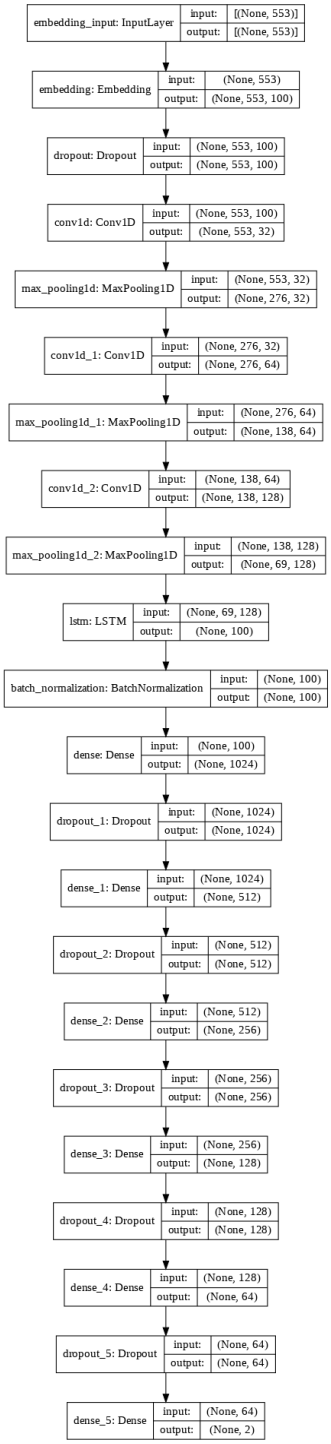


Figure 5.2: LSTM + CNN architecture

5.1.1 Results

The best validation loss was obtained, at the 4th epoch out of the total 20. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 5.1. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 5.3

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
4	0.8717	0.3057	0.8551	0.3411	0.8525

Table 5.1: LSTM + CNN: Classification scores.

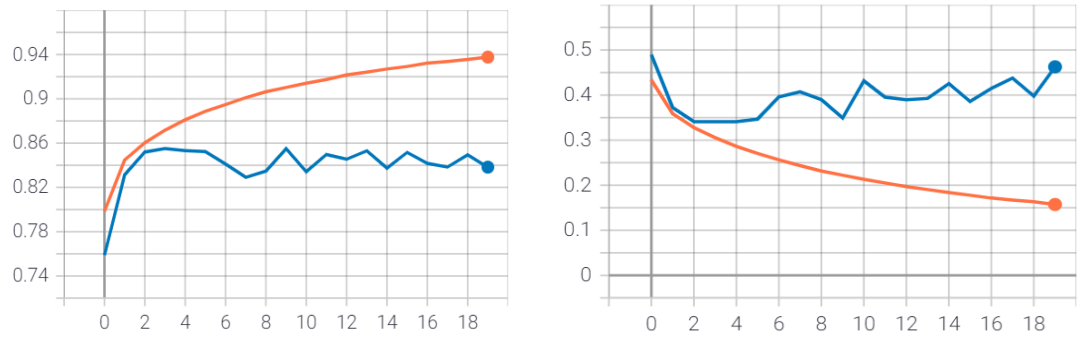


Figure 5.3: LSTM + CNN: Training and Validation graphs.
Left: accuracy vs epoch.
Right: loss vs epoch.

5.2 BiGRU + CapsuleNet

The first layer is the Embedded layer, which represents each word with 300 length vectors. I use word embeddings from pre-trained Glove and paragram. These embedding were combined using meta-embedding. The following layer is the Bidirectional GRU layer, which has 100 memory units. A capsule network with 10 units are added after this layer. The features are flattened and are passed through Dense layers along with Dropout and batch normalization layers. Finally, we use a Dense output layer with a 2 neuron and a softmax activation function.

Trained the model for 20 epochs with batch size 1024, Adam optimizer, binary

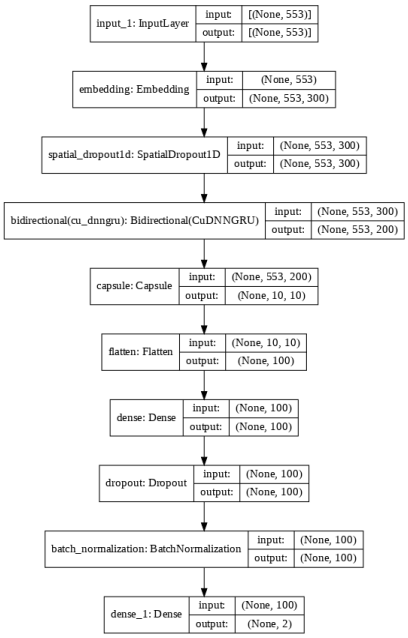


Figure 5.4: BiGRU + CapsuleNet architecture

crossentropy loss and 3 callback functions – [CSV logger, tensorboard, Model Checkpoint]

5.2.1 Results

The best validation loss was obtained, at the 3rd epoch out of the total 20. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 5.2. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 5.5

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
3	0.9105	0.2192	0.8598	0.3572	0.8618

Table 5.2: BiGRU + CapsuleNet: Classification scores.

5.3 BiLSTM + BiGRU + attention

The first layer is the Embedded layer, which represents each word with 600 length vectors. I used word embeddings from pre-trained Glove and fasttext. These embedding were concatenated together. The following layer is the Bidirectional LSTM layer, which

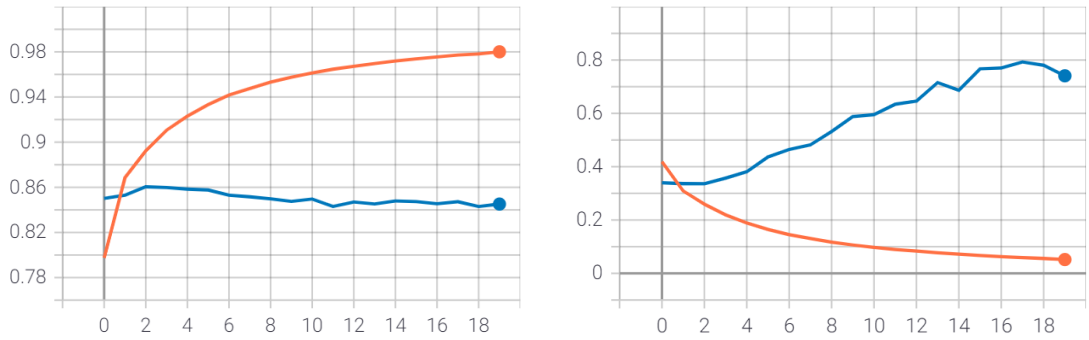


Figure 5.5: BiGRU + CapsuleNet: Training and Validation graphs.
Left: accuracy vs epoch.
Right: loss vs epoch.

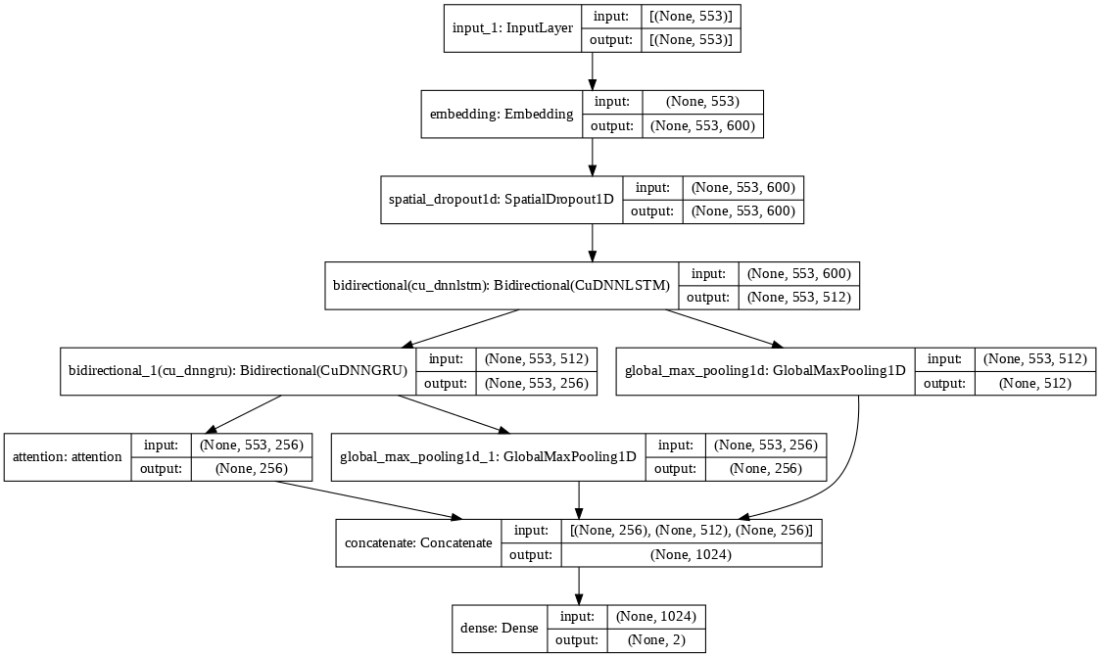


Figure 5.6: BiLSTM + BiGRU + attention architecture

has 100 memory units. Followed by Bidirectional GRU layer with 100 memory units. A attention layer is added after the BiGRU. The output of BiGRU and BiLSTM are passed through the Global Max pooling layer. All the features are concatenated. Finally, we use a Dense output layer with a 2 neuron and a softmax activation function.

Trained the model for 15 epochs with batch size 512, Adam optimizer, binary crossentropy loss and 3 callback functions – [CSV logger, tensorboard, Model Checkpoint]

5.3.1 Results

The best validation loss was obtained, at the 6th epoch out of the total 15. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 5.3. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 5.7

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
6	0.9090	0.2242	0.8789	0.3021	0.8790

Table 5.3: BiLSTM + BiGRU + attention: Classification scores.

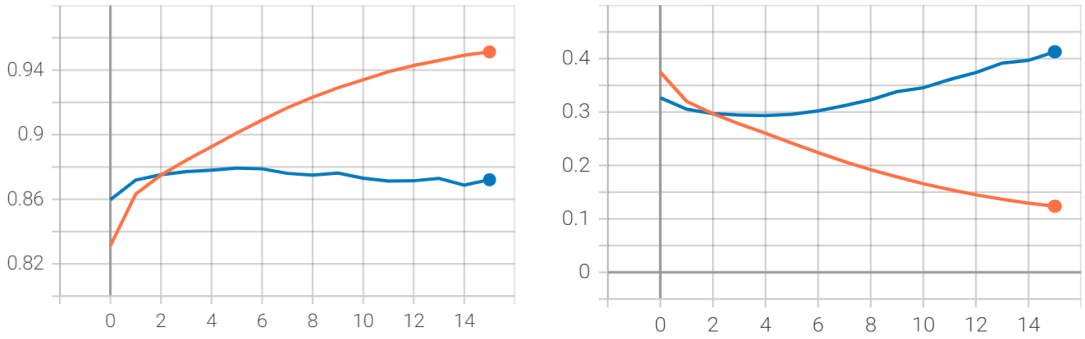


Figure 5.7: BiLSTM + BiGRU + attention: Training and Validation graphs.
Left: accuracy vs epoch.
Right: loss vs epoch.

5.4 2D CNN

The first layer is the Embedded layer, which represents each word with 600 length vectors I used word embeddings from pre-trained Glove and fasttext. These embedding were concatenated together. The word embeddings are reshaped and 2D convolution is applied with different filter sizes. (refer Fig. 5.9). Followed by Max Pooling layer. The outputs of all the max pooling operations are concatenated. Finally, we use a Dense output layer with a 2 neuron and a softmax activation function

Trained the model for 15 epochs with batch size 1024, Adam optimizer, binary crossentropy loss and 3 callback functions – [CSV logger, tensorboard, Model Checkpoint]

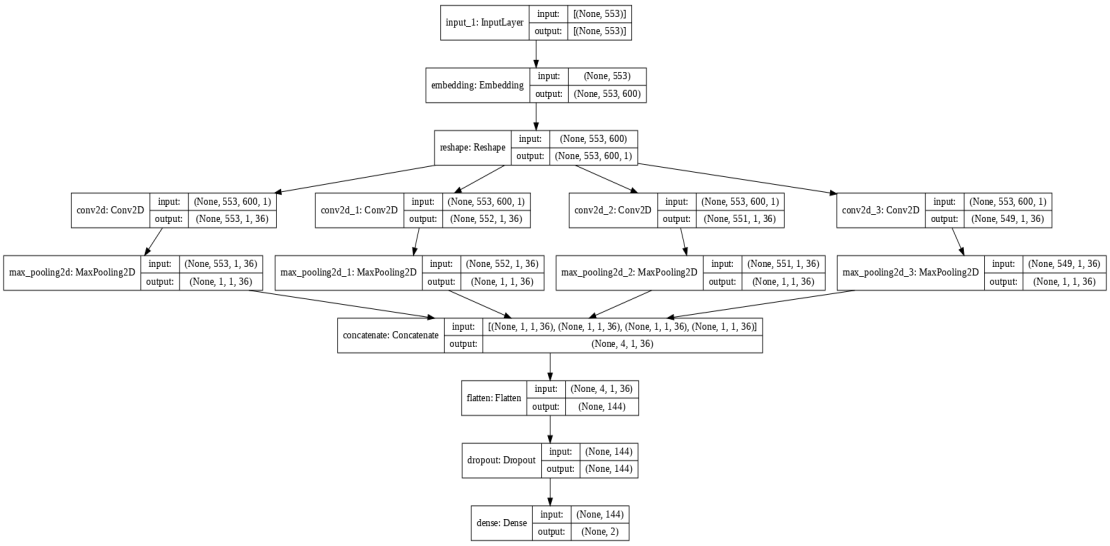


Figure 5.8: 2D CNN architecture

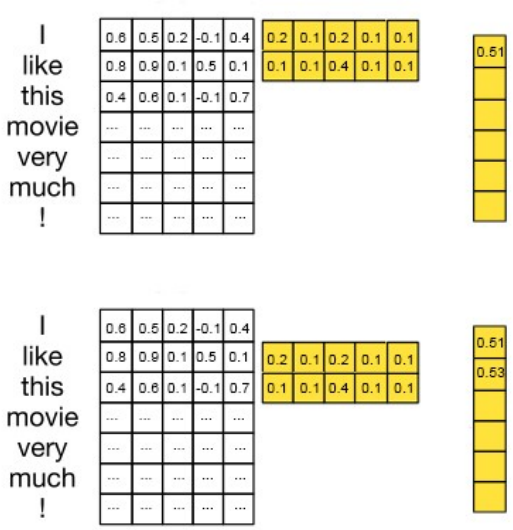


Figure 5.9: 2D convolution on Word embedding matrix

5.4.1 Results

The best validation loss was obtained, at the 2nd epoch out of the total 15. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 5.4. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 5.10

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
2	0.8906	0.267	0.8652	0.3283	0.8677

Table 5.4: 2D CNN: Classification scores.

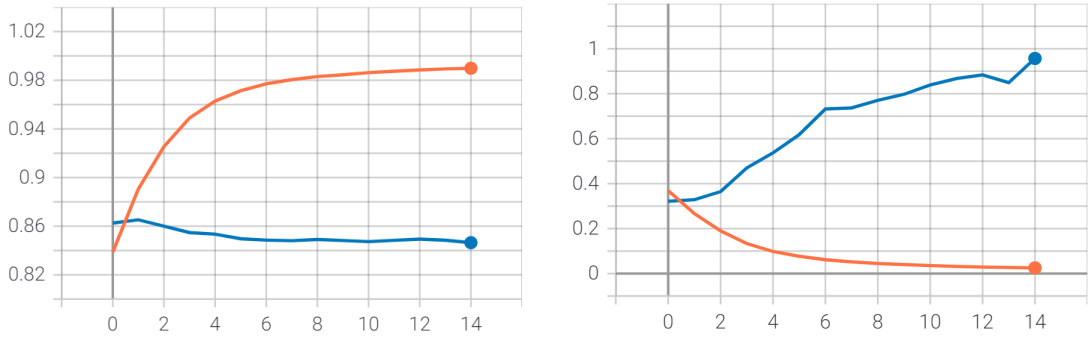


Figure 5.10: 2D CNN: Training and Validation graphs.
Left: accuracy vs epoch.
Right: loss vs epoch.

5.5 BERT + Dense

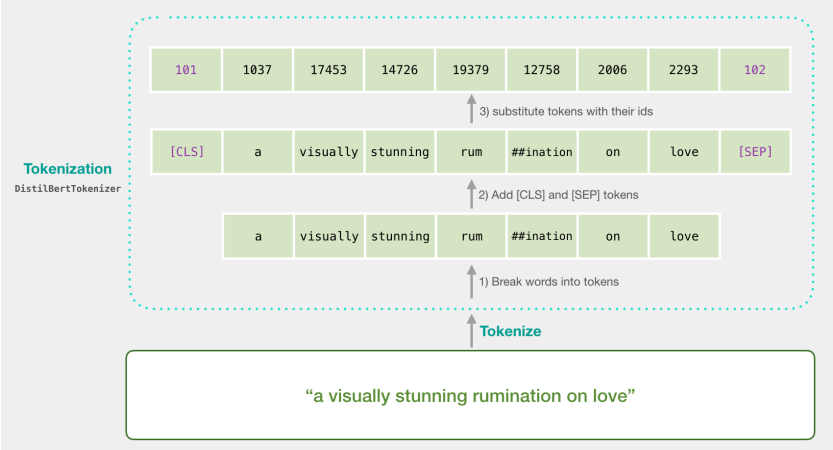


Figure 5.11: Text Pre-Processing for BERT

The first layer is a input layer which takes pre-processed text (refer Figure. 5.11) as the input and passes it through the BERT model. There are two outputs: a pooled_output of shape [batch_size, 768] with representations for the entire input sequences and a sequence_output of shape [batch_size, max_seq_length, 768] with representations for each input token (in context). We will slice the sequence_output to take the 768 dimension contextual embedding and pass it through the 2 dense layers of 64 and 32 units. Finally, we use a Dense output layer with a 2 neuron and a softmax activation function.

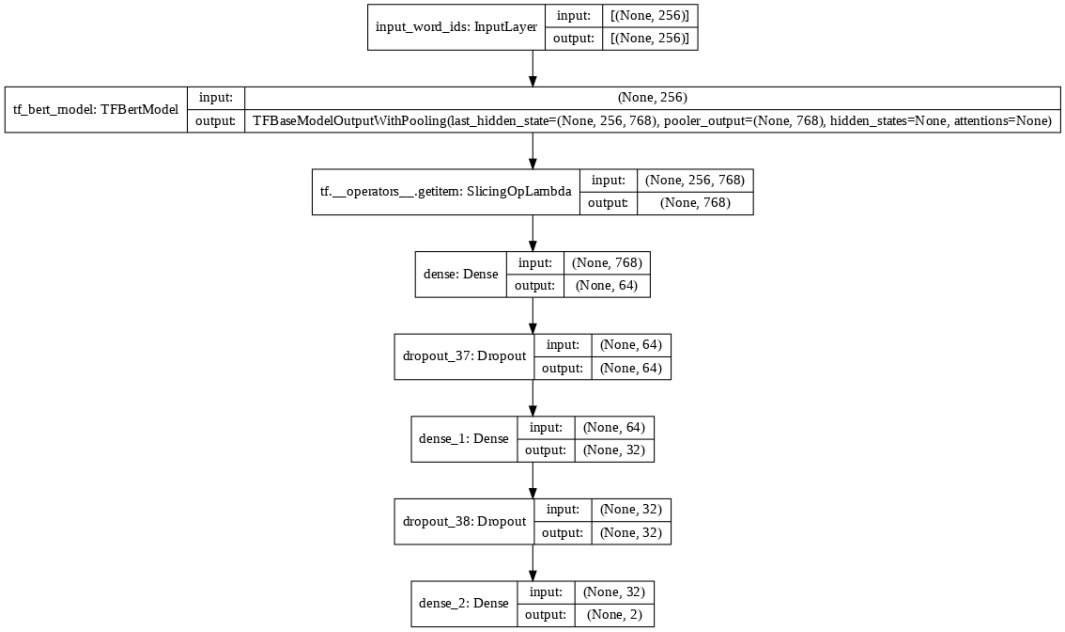


Figure 5.12: BERT + Dense architecture

Fine tuned the model for 10 epochs with batch size 144, Adam optimizer, binary crossentropy loss and 1 callback functions – [Model Checkpoint]

5.5.1 Results

The best validation loss was obtained, at the 3rd epoch out of the total 10. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 5.5. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 5.13

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
3	0.9049	0.2413	0.8934	0.2726	0.8946

Table 5.5: BERT + Dense: Classification scores.

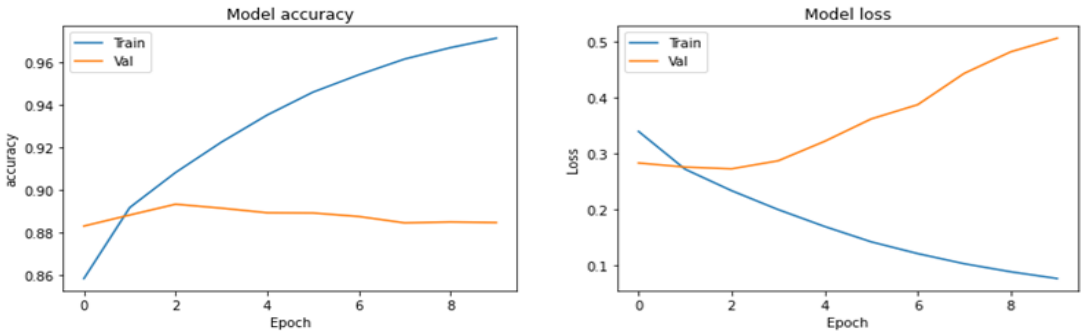


Figure 5.13: BERT + Dense: Training and Validation graphs.

5.6 RoBERTa + Dense

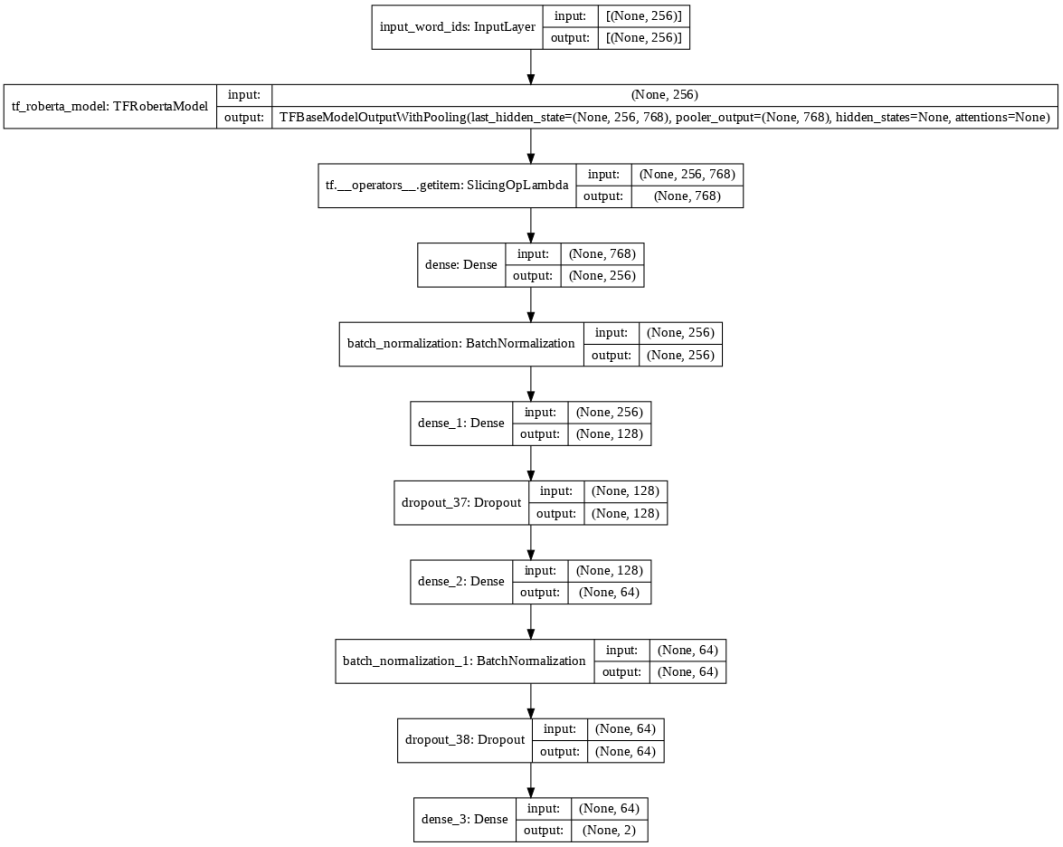


Figure 5.14: RoBERTa + Dense architecture

The first layer is a input layer which takes pre-processed text (refer Figure. 5.11) as the input and passes it through the RoBERTa model. There are two outputs: a pooled_output of shape [batch_size, 768] with representations for the entire input sequences and a sequence_output of shape [batch_size, max_seq_length, 768] with representations for each input token (in context). We will slice the sequence_output to take

the 768 dimension contextual embedding and pass it through the various Dense layers along with Dropout and batch normalization layers. Finally, we use a Dense output layer with a 2 neuron and a softmax activation function.

Fine tuned the model for 10 epochs with batch size 144, Adam optimizer, binary crossentropy loss and 1 callback functions – [Model Checkpoint]

5.6.1 Results

The best validation loss was obtained, at the 6th epoch out of the total 10. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 5.6. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 5.15

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
6	0.9040	0.2555	0.8852	0.3061	0.8862

Table 5.6: RoBERTa + Dense: Classification scores.

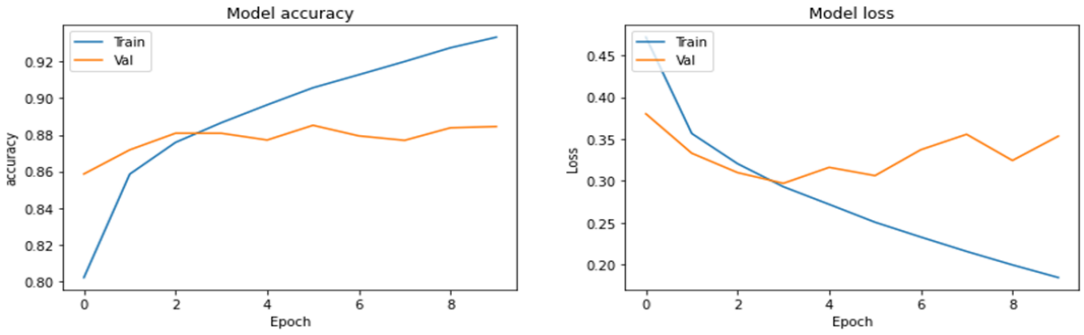


Figure 5.15: RoBERTa + Dense: Training and Validation graphs.

Performance Comparison

Table 5.7 displays both the baselines results and our text modality models result on Fakeddit dataset. We report our model's validation and test accuracy. We can easily see how much better our text models does than the baseline methods.

Text model	**Validation accuracy**	**Test accuracy**
BERT (Baseline)	0.8654	0.8644
InferSent (Baseline)	0.8634	0.8631
LSTM + CNN	0.8551	0.8525
BiGRU + Capsule	0.8598	0.8618
BiLSTM + BiGRU + attention	0.8789	0.8790
2D CNN	0.8652	0.8677
BERT + Dense	**0.8934**	**0.8946**
RoBERTa + Dense	0.8852	0.8862

Table 5.7: Performance of Text modality models with the Baselines

CHAPTER 6

Fake News Detection based on Image Manipulation

In this chapter, we implemented and evaluated various image modality models. The models were trained on the Fakeddit dataset and used only the visual information (images) in posts to determine whether they were manipulated (fake) or not.

As seen previously, Deep Neural Networks can be used to detect image manipulation, without the need for a feature extraction task. Both feature extraction and classification can be done in an end-to-end process. Convolution Neural Networks (CNNs) have been shown to be extremely efficient in detecting image manipulation.

6.1 Fine-Tuning Inception-ResNet-v2

The convolutional neural network Inception-ResNet-v2 was trained on over a million images from the ImageNet database. The 164-layer network can classify images into 1000 different classes. As a result, the network has learned a variety of rich feature representations for a variety of images.

It's built on a foundation of the Inception framework and the Residual connection. Multiple sized convolutional filters are combined with residual connections in the Inception-Resnet block. (refer Fig. 6.1, 6.2) The use of residual connections not only prevents the vanishing gradient issue caused by deep structures, but it also cuts training time in half.

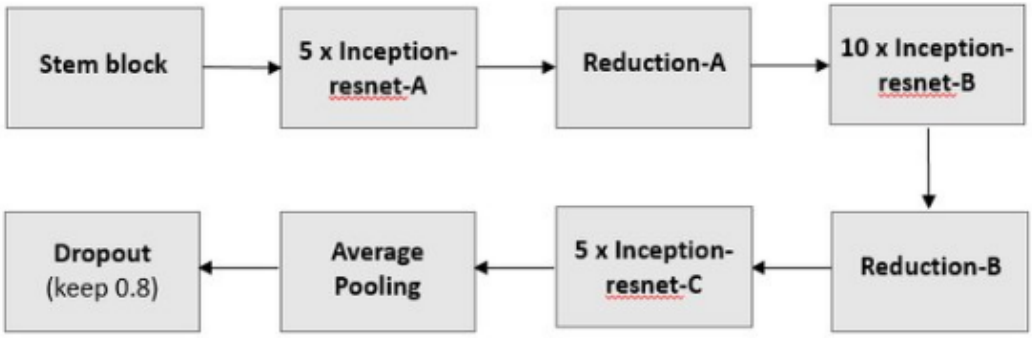


Figure 6.1: Inception-Resnet-v2 architecture. [4]

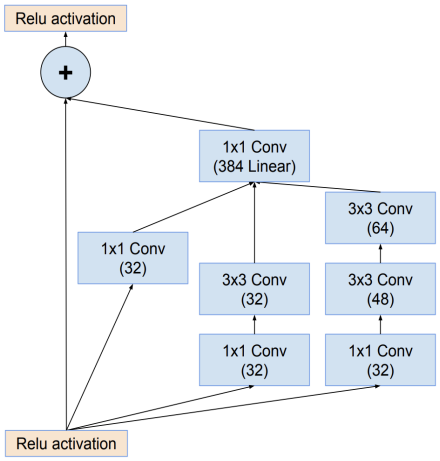


Figure 6.2: Residual Inception Block(Inception-ResNet-A). [4]

6.1.1 Implementation

We fine-tuned Inception-ResNet-v2 network on the Fakeddit dataset with the following parameters:

- The convolutional part of the model is instantiated and pre-trained weights from ImageNet are loaded. The fully connected classifer with softmax predictor/activation function is added on top of the convolutional part. The design of the model is depicted in Figure 6.3.
- The entire model was unfrozen, i.e, it was made trainable and retrained on the Fakeddit dataset.
- Image rescaled to 150x150 pixels.
- Batch size = 256.

- Adam Optimizer used with learning rate = 0.0005
- The model is trained for 15 epochs.

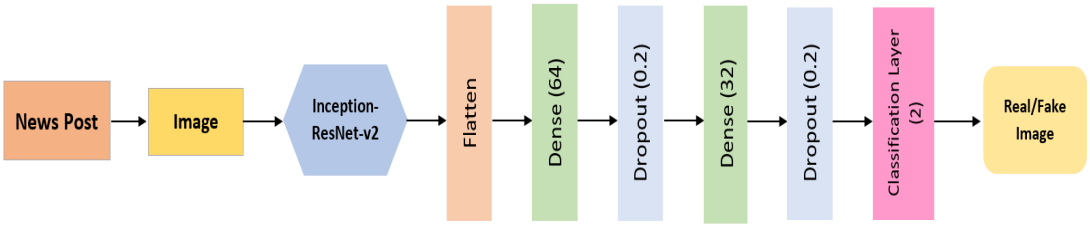


Figure 6.3: Inception-ResNet-v2 model design

6.1.2 Results

The best validation loss was obtained, at the 5th epoch out of the total 15. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 6.1. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 6.4

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
5	0.8505	0.3276	0.8049	0.6874	0.8066

Table 6.1: Inception-ResNet-v2: Classification scores.

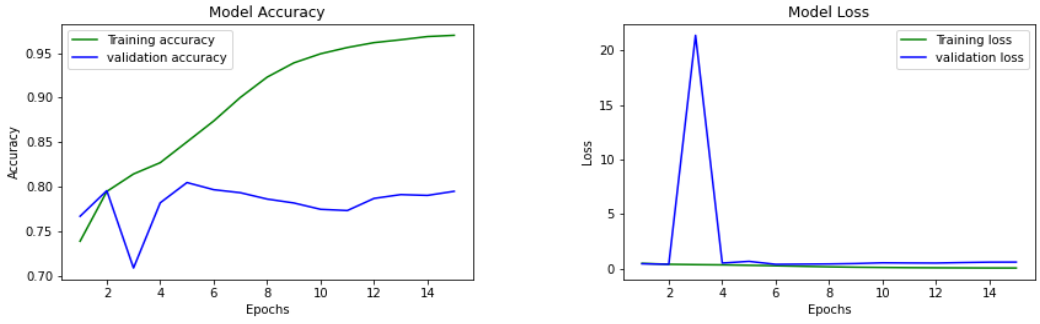


Figure 6.4: Inception-ResNet-v2: Training and Validation graphs.

The confusion matrix and classification report are presented in Figure 6.5.

	precision	recall	f1-score	support
Fake	0.86	0.81	0.83	35763
Real	0.73	0.81	0.77	23507
accuracy			0.81	59270
macro avg	0.80	0.81	0.80	59270
weighted avg	0.81	0.81	0.81	59270

Figure 6.5: Inception-ResNet-v2: Confusion matrix and classification report

6.2 Fine-Tuning Xception Model

Francois Chollet proposed the Xception Model. Xception is an extension of the Inception Architecture that uses depthwise Separable Convolutions to replace the regular Inception modules.

The data first passes through the entry flow, then eight times through the middle flow, and finally through the exit flow. Batch normalisation is applied on both Convolution and SeparableConvolution layers. (Refer Fig. 6.6 for Xception network architecture.)

In most standard classification problems, the Xception architecture outperformed VGG-16, ResNet, and Inception V3.

6.2.1 Implementation

We fine-tuned Xception network on the Fakeddit dataset with the following parameters:

- The convolutional part of the model is instantiated and pre-trained weights from ImageNet are loaded. The fully connected classifer with softmax predictor/activation function is added on top of the convolutional part. The design of the model is depicted in Figure 6.7.
- The entire model was unfrozen, i.e, it was made trainable and retrained on the Fakeddit dataset.
- Image rescaled to 150x150 pixels.
- Batch size = 256.
- Adam Optimizer used with learning rate = 0.0005

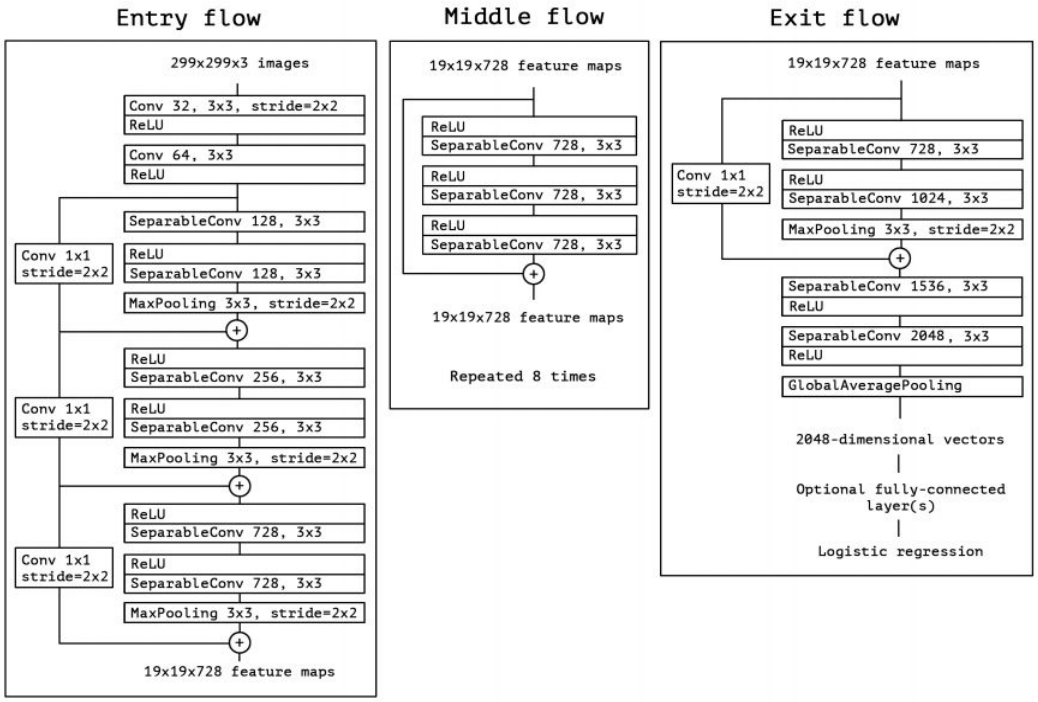


Figure 6.6: Xception architecture. [5]

- The model is trained for 15 epochs.

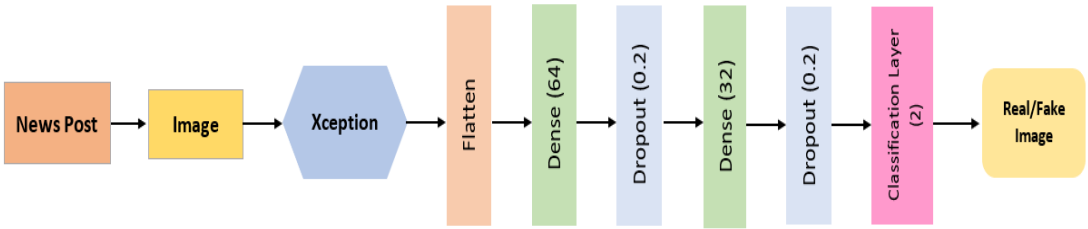


Figure 6.7: Xception model design

6.2.2 Results

The best validation loss was obtained, in the 2nd epoch. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 6.2. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 6.8

The confusion matrix and classification report are presented in Figure 6.9. It is observed that the recall for fake images is low as compared to previous model.

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
2	0.8214	0.3780	0.8207	0.3751	0.8232

Table 6.2: Xception network: Classification scores.

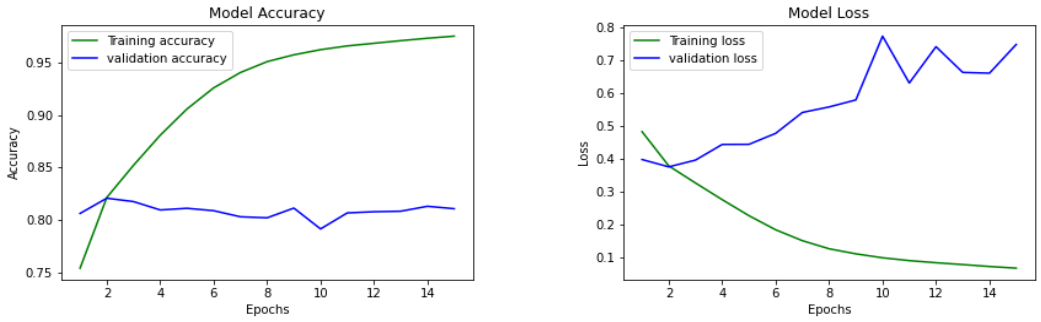


Figure 6.8: Xception: Training and Validation graphs.

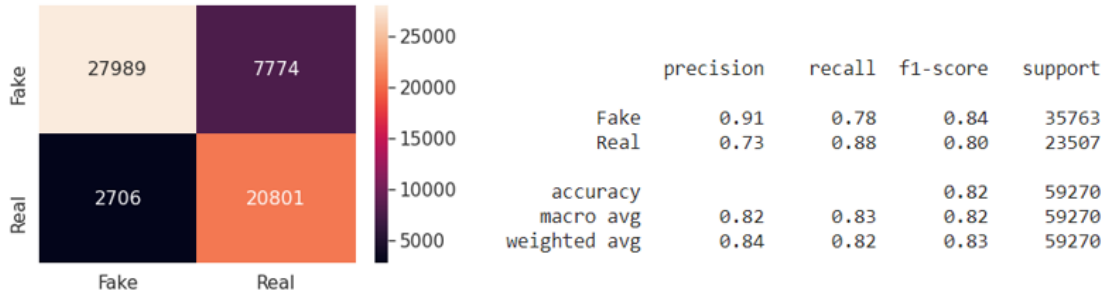
6.3 Fine-tuning Approach based on Error Level Analysis

In the previous sections, two models were fine-tuned on Fakeddit dataset images. While overall F1-scores improved, increasing recall scores for fake images while also maintaining good precision still remained as a challenge.

6.3.1 Transforming images

In the pursuit of improving the performance of the model and increasing the recall of fake images, a change of CNN architecture though useful, did not bring about significant improvements. One way identified to deal with this issue is to transform the input images to the model in a way which aids in picking up distinguishing features between fake and real images. This can help improve the recall scores of fake images.

For this purpose, Error Level Analysis was explored.



	precision	recall	f1-score	support
Fake	0.91	0.78	0.84	35763
Real	0.73	0.88	0.80	23507
accuracy			0.82	59270
macro avg	0.82	0.83	0.82	59270
weighted avg	0.84	0.82	0.83	59270

Figure 6.9: Xception: Confusion matrix and classification report

6.3.2 The Method

Error Level Analysis (ELA) is a forensic technique which analyses compression artifacts in images which are compressed with lossy techniques such as JPEG. It helps identifying regions in the image which have different compression levels.

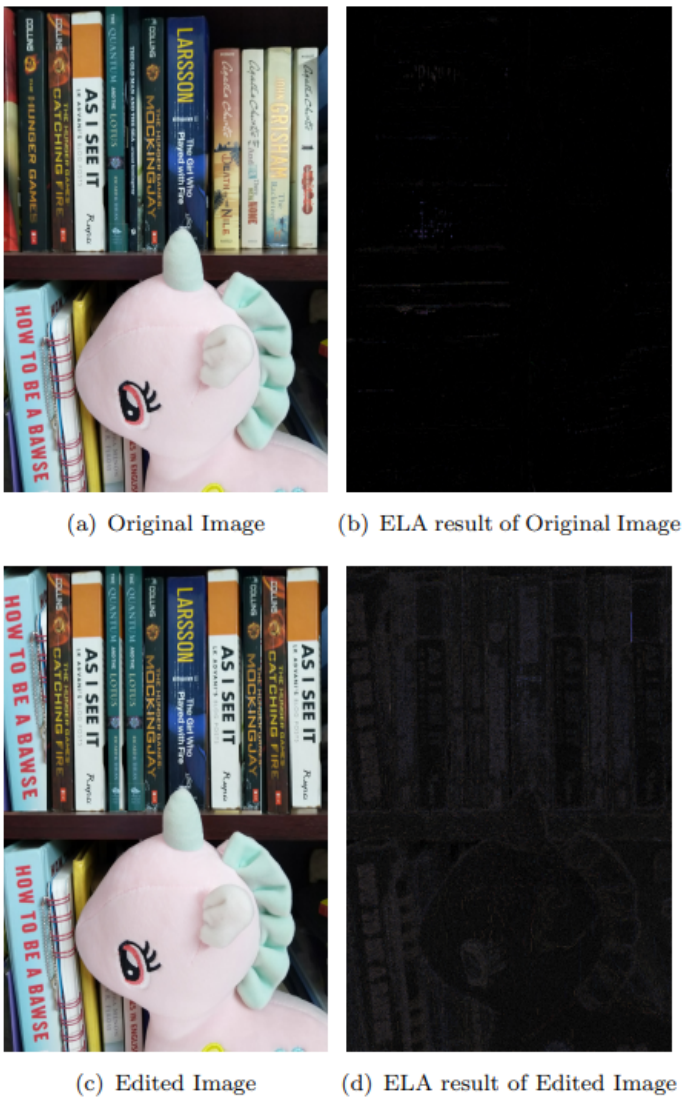
Usually, the entire picture has about the same compression level, since lossy compression (such as JPEG) is applied uniformly over the image. However, certain sections of the image may have significantly different error levels, because those sections may have been subjected to the same lossy compression different number of times, or a different type of lossy compression. Therefore, a difference in error level in different sections of an image that indicates that it is likely that the image has been edited.

JPEG compression splits an image into blocks of 8x8 pixels and transformations such as Discrete Cosine Transform (DCT) are applied block-wise. According to [37], the amount of error introduced by each resave of the image is not linear, i.e., saving an image at say 70% quality and then resaving to 90% generates the same image as a 90% resaved to 70%, or an image saved one-time at 63%. The amount of error is limited to the 8x8 cells and after about 64 resaves, there is virtually no change. when an image is modified, the 8x8 cells containing the modification are no longer at the same error level as the rest of the unmodified image. ELA works by first intentionally re-saving the image at a known error level (e.g. at 90%) and then computing the difference between the images. With each resave, the error level potential is lowered, resulting in a 'darker' ELA result (shall be illustrated below). If there is no change, that means that the cell has reached the local minima for error at that quality level. If the picture is modified, then the regions which had no additional error (stable) become unstable because of those

alterations. [37]

6.3.3 Demonstrating Error Level Analysis

Figure 6.10(a) shows an original, unedited image captured by a smartphone camera. The figure is resaved at 95% quality and the Error Level Analysis in Figure 6.10(b) shows that most colors have been compressed well. (Darker ELA values mean lower error levels) Figure 6.10(c) shows an altered version of Figure 6.10(a) - some books have been copied. The ELA Analysis (Figure 6.10(d)) shows higher error levels at tampered regions.



(a) Original Image (b) ELA result of Original Image

(c) Edited Image (d) ELA result of Edited Image

Figure 6.10: ELA Analysis of Original and Edited Images.

6.3.4 Computing ELA images

As seen previously, Error Level Analysis(ELA) helps identify digitally altered images since the error levels throughout such images are not uniform. Therefore, the next step is to compute ELA images for all the images in the Fakeddit dataset and use this for fine-tuning CNN architectures. This is done by resaving all the images at 90% quality and calculating the difference between the image and its resaved version.

6.4 Fine-Tuning Inception-ResNet-v2 with ELA images

We fine-tuned Inception-ResNet-v2 network on the computed ELA images with the following parameters:

- The convolutional part of the model is instantiated and pre-trained weights from ImageNet are loaded. The fully connected classifer with softmax predictor/activation function is added on top of the convolutional part. The design of the model is depicted in Figure 6.11.
- The entire model was unfrozen, i.e, it was made trainable and retrained on the ELA version of Fakeddit dataset.
- Image rescaled to 150x150 pixels.
- Batch size = 256.
- Adam Optimizer used with learning rate = 0.0005
- The model is trained for 15 epochs.

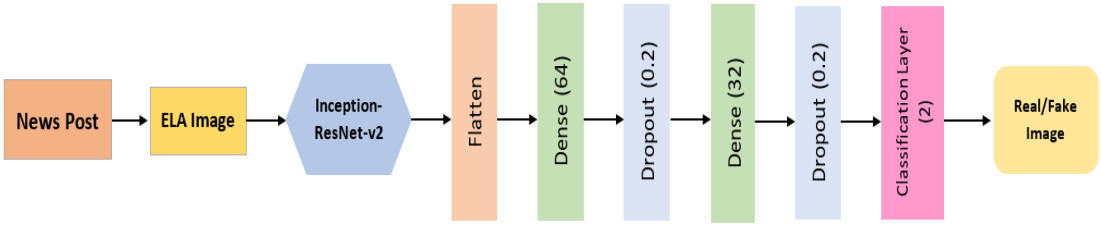


Figure 6.11: Inception-ResNet-v2 with ELA model design

6.4.1 Results

The best validation loss was obtained, at the 2nd epoch. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 6.3. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 6.12

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
2	0.7938	0.4074	0.7952	0.3979	0.7970

Table 6.3: Inception-ResNet-v2 with ELA: Classification scores.

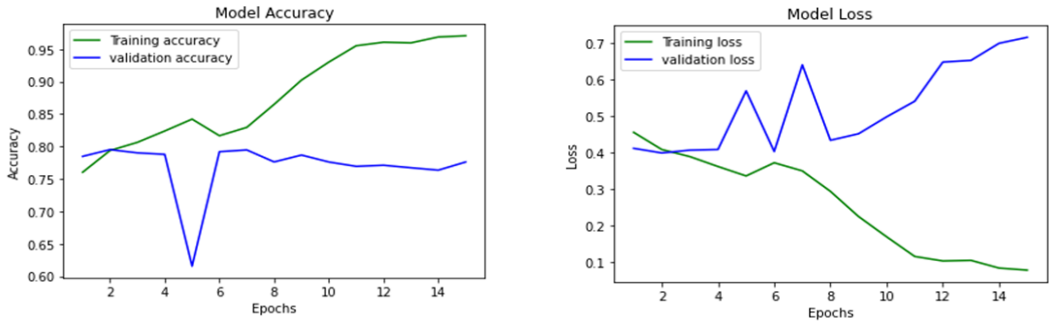
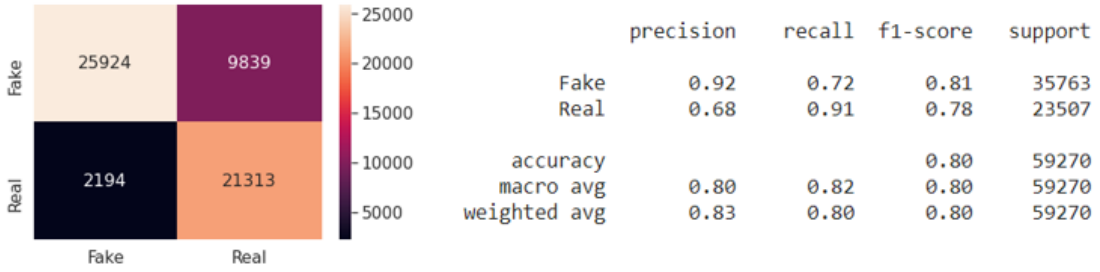


Figure 6.12: Inception-ResNet-v2 with ELA: Training and Validation graphs.

The confusion matrix and classification report are presented in Figure 6.13.



	precision	recall	f1-score	support
Fake	0.92	0.72	0.81	35763
Real	0.68	0.91	0.78	23507
accuracy			0.80	59270
macro avg	0.80	0.82	0.80	59270
weighted avg	0.83	0.80	0.80	59270

Figure 6.13: Inception-ResNet-v2 with ELA: Confusion matrix and classification report

6.5 Fine-Tuning ResNet50 with ELA images

ResNet, short for Residual Networks, is a well-known neural network that serves as the foundation for many computer vision tasks. In 2015, this model was the winner

of the ImageNet competition. ResNet was a game-changer because it allowed us to successfully train extremely deep neural networks with 150+ layers. Due to the issue of vanishing gradients, training very deep neural networks was difficult before ResNet. ResNet50 is a ResNet variant with 48 Convolution layers, 1 MaxPool layer, and 1 Average Pool layer. It has a total of 3.8 x 10^9 floating-point operations.

Implementation

We fine-tuned ResNet50 network on the computed ELA images with the following parameters:

- The convolutional part of the model is instantiated and pre-trained weights from ImageNet are loaded. The fully connected classifer with softmax predictor/activation function is added on top of the convolutional part. The design of the model is depicted in Figure 6.14.
- The entire model was unfrozen, i.e, it was made trainable and retrained on the ELA version Fakeddit dataset.
- Image rescaled to 150x150 pixels.
- Batch size = 256.
- Adam Optimizer used with learning rate = 0.0005
- The model is trained for 15 epochs.

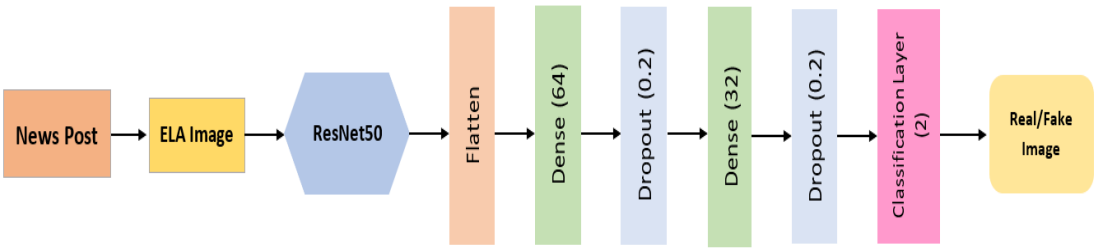


Figure 6.14: ResNet50 with ELA model design

6.5.1 Results

The best validation loss was obtained, at the 4th epoch. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
4	0.7939	0.4032	0.7901	0.4046	0.7958

Table 6.4: ResNet50 with ELA: Classification scores.

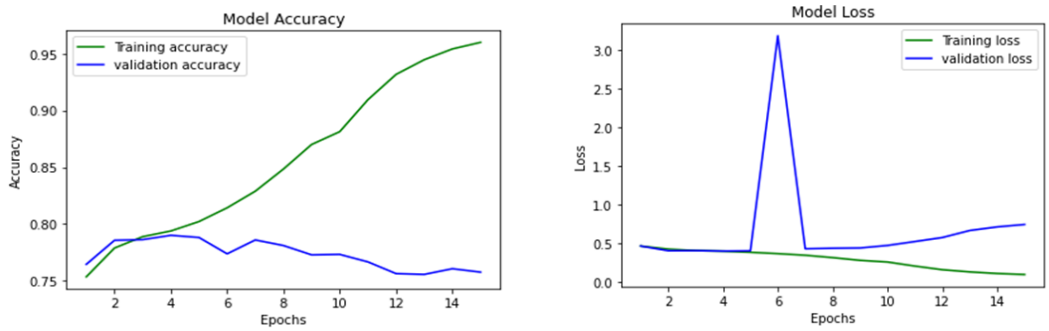


Figure 6.15: ResNet50 with ELA: Training and Validation graphs.

are presented in Table 6.4. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 6.15

The confusion matrix and classification report are presented in Figure 6.16.

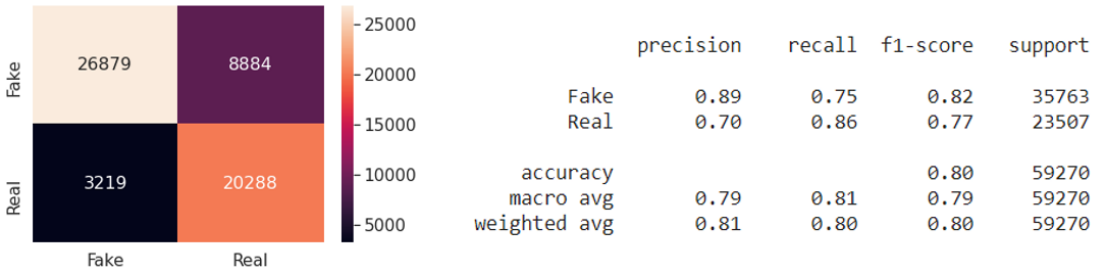


Figure 6.16: ResNet50 with ELA: Confusion matrix and classification report

6.6 Fine-Tuning Xception Model with ELA images

We fine-tuned Xception network on the computed ELA images with the following parameters:

- The convolutional part of the model is instantiated and pre-trained weights from ImageNet are loaded. The fully connected classifer with softmax predictor/activation function is added on top of the convolutional part. The design of the model is depicted in Figure 6.17.
- The entire model was unfrozen, i.e, it was made trainable and retrained on the ELA version of Fakeddit dataset.

- Image rescaled to 150x150 pixels.
- Batch size = 256.
- Adam Optimizer used with learning rate = 0.0005
- The model is trained for 10 epochs.

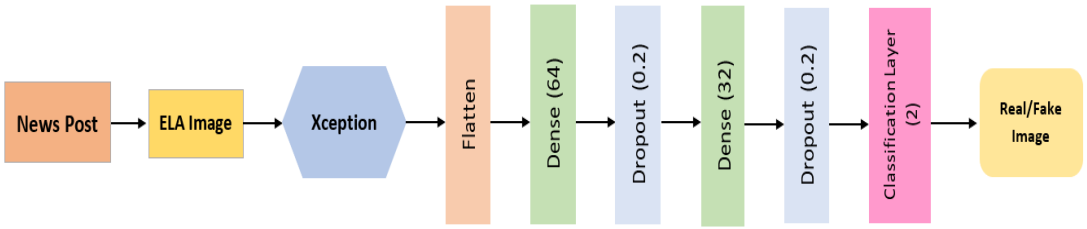


Figure 6.17: Xception with ELA model design

6.6.1 Results

The best validation loss was obtained, in the 2nd epoch. The weights obtained at this point are used to evaluate the model against the test dataset. The classification scores are presented in Table 6.5. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 6.18

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
2	0.8032	0.3922	0.7958	0.3981	0.8001

Table 6.5: Xception network with ELA: Classification scores.

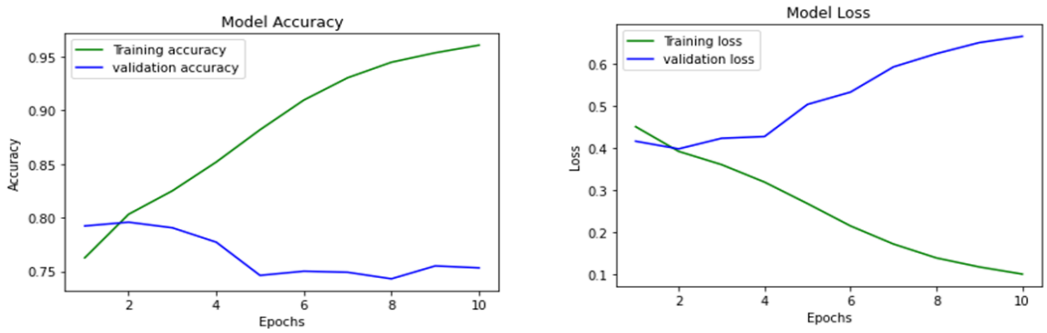
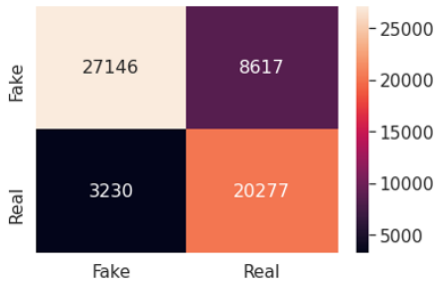


Figure 6.18: Xception with ELA: Training and Validation graphs.

The confusion matrix and classification report are presented in Figure 6.19.



	precision	recall	f1-score	support
Fake	0.89	0.76	0.82	35763
Real	0.70	0.86	0.77	23507
accuracy			0.80	59270
macro avg	0.80	0.81	0.80	59270
weighted avg	0.82	0.80	0.80	59270

Figure 6.19: Xception with ELA: Confusion matrix and classification report

Image models	Validation accuracy	Test accuracy
VGG16 (Baseline)	0.7355	0.7376
EfficientNet (Baseline)	0.6115	0.6087
ResNet50 (Baseline)	0.8043	0.8070
Inception-ResNet-v2	0.8049	0.8066
Xception	**0.8207**	**0.8232**
Inception-ResNet-v2 with ELA images	0.7952	0.7970
ResNet50 with ELA images	0.7901	0.7958
Xception with ELA images	0.7958	0.8001

Table 6.6: Performance of Image modality models with the Baselines

6.7 Performance Comparison

Table 6.6 displays both the baselines results and our image modality models result on Fakeddit dataset. We report our model's validation and test accuracy. We can see how much better our proposed method performs compared to the baseline methods. Models that have been fine-tuned with ELA images have lower validation and test accuracy than models that have been fine-tuned with normal images. This could be because the Fakeddit dataset contains a lower proportion of tampered/manipulated images.

CHAPTER 7

Fake News Detection based on Image Caption & Image Manipulation

In the previous chapters,We found that single-modality approaches yielded promising results, but the unstructured nature of social networking data makes information extraction difficult. Using either text or an image alone may not be enough to detect falsification. For example, the text in Fig. 7.1 states, "the presence of sharks during Hurricane Sandy 2012," A closer look at the picture shows that it was spliced together to include fake sharks. In this case, textual claims were insufficient to detect falsification.

Figure 7.1: During Hurricane Sandy in 2012, photos of spliced sharks were taken [6]

In this chapter, we implemented and evaluated various multi-modality (i.e. text and image) models to overcome this stumbling block. The models were trained on the Fakeddit dataset and used both the textual (image captions) and the visual information (images) in posts to determine whether they were fake or real.

7.1 Fine-tuning multimodal (text + image) models

In Chapter 5, we fine-tuned the BERT model using just the textual information i.e., the image captions present in the post were used for fine-tuning the BERT model.

Similar in Chapter 6, the Xception model was fine-tuned using the image present in the post.

To use the power of both techniques, an ensemble model was designed to improve the identification of fake news - with Xception network to help in identifying images with high digital alterations, and BERT to make use of the potential of language models i.e., the model will learn contextual information. Contextual information is essential in addition to content information since real-world texts, images, and videos are complex.

The design of the ensemble construction is depicted in Figure 7.2 The layers from the Xception model fine-tuned on Fakeddit dataset images make up the rightmost vertical branch, while the layers from the BERT model fine-tuned on image captions make up the leftmost vertical branch. All layers from these models except the last classification layers are included in this ensemble model. The final module is a multimodal fusion module, which incorporates representations obtained from various modalities. (i.e. text and image) together to form news feature vector. For fake news classification, this news representation is fed into a fully connected neural network with softmax activation.

The ensemble model is loaded with the best weights obtained from 2 models which were trained independently in previous chapters - fine-tuning Xception on Fakeddit dataset images and fine-tuning BERT on image captions(i.e. text). The ensemble model is fine-tuned again on Fakeddit dataset samples that have both image and text. The following phases detail some experiments performed to obtain optimal weights.

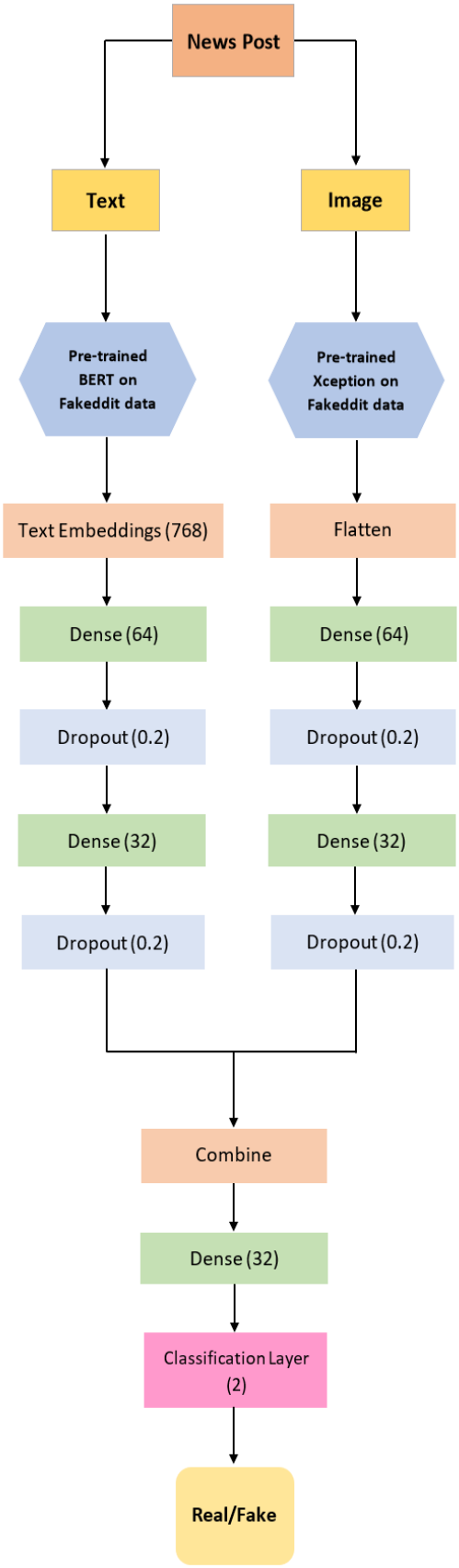


Figure 7.2: Multimodal (text + image) Model Design

7.1.1 Phase 1

- All layers in the Xception branch are made untrainable until the Flatten layer, and all layers in the BERT branch are made untrainable until the Text embeddings. In other words, we made the base models in both branches untrainable. We will train only the fully connected dense layers.
- Both modalities' 32-dimensional vectors are combined using **maximum fusion method** and fed into a fully linked neural network classifier with a 32-layer hidden layer and a 2-layer classification layer with softmax activation. (Refer Fig. 7.3)
- Batch size = 256
- Adam optimiser is set with a learning rate of 0.0005 and the model is trained for 15 epochs.

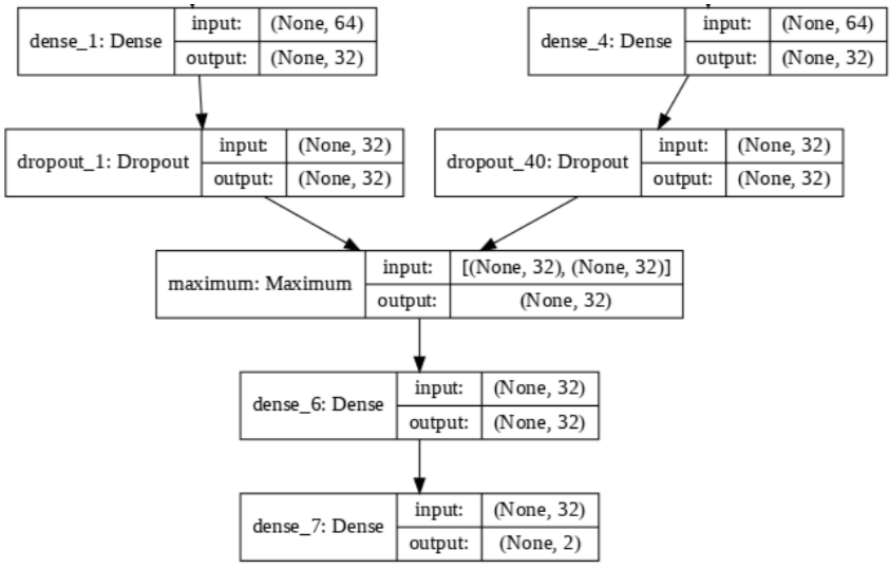


Figure 7.3: Multimodal (text + image) Model Snippet with Max fusion method.

Results

The best validation loss was obtained, at the 7th epoch out of a total of 15 epochs.The model is evaluated against the test dataset using the weights obtained at this stage. The classification scores are presented in Table 7.1. The loss variations and accuracy over the epochs for training and validation sets are shown in Figure 7.4

The confusion matrix and classification report are presented in Figure 7.5. It is observed that the accuracy, precision, recall and F1-score have significantly improved as compared to the unimodal models.

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
7	0.9438	0.1393	0.9161	0.2188	0.9187

Table 7.1: Multimodal Model with Max fusion method: Classification scores.

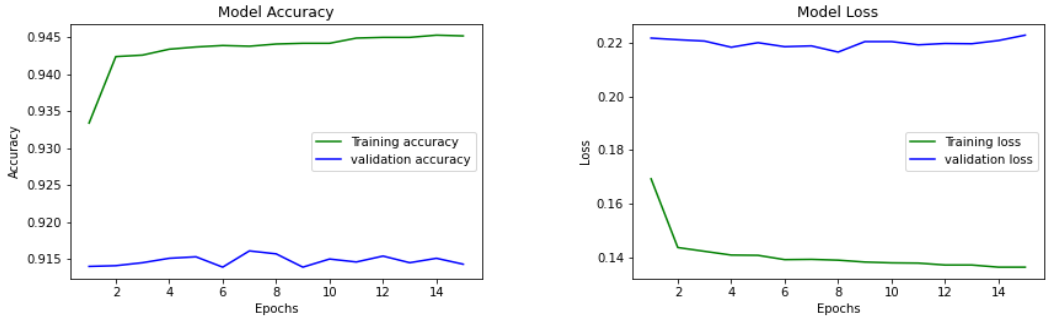


Figure 7.4: Multimodal Model with Max fusion method: Training and Validation graphs.

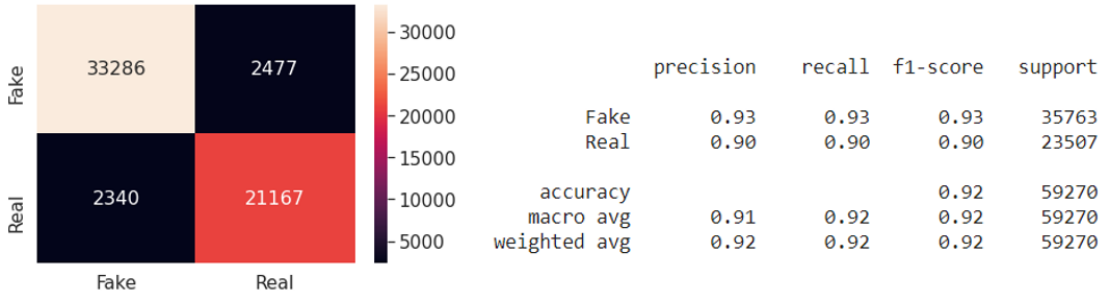


Figure 7.5: Multimodal Model with Max fusion method: Confusion matrix and classification report

7.1.2 Phase 2

For this phase, the following changes are made:

- Both modalities' 32-dimensional vectors are combined using **concatenation fusion method** and fed into a fully linked neural network classifier with a 32-layer hidden layer and a 2-layer classification layer with softmax activation. (Refer Fig. 7.6)
- No changes made in optimizer settings.

Results

The best validation loss was obtained, at the 8th epoch out of a total of 15 epochs. The weights obtained at this point are used to evaluate the model against the test dataset.

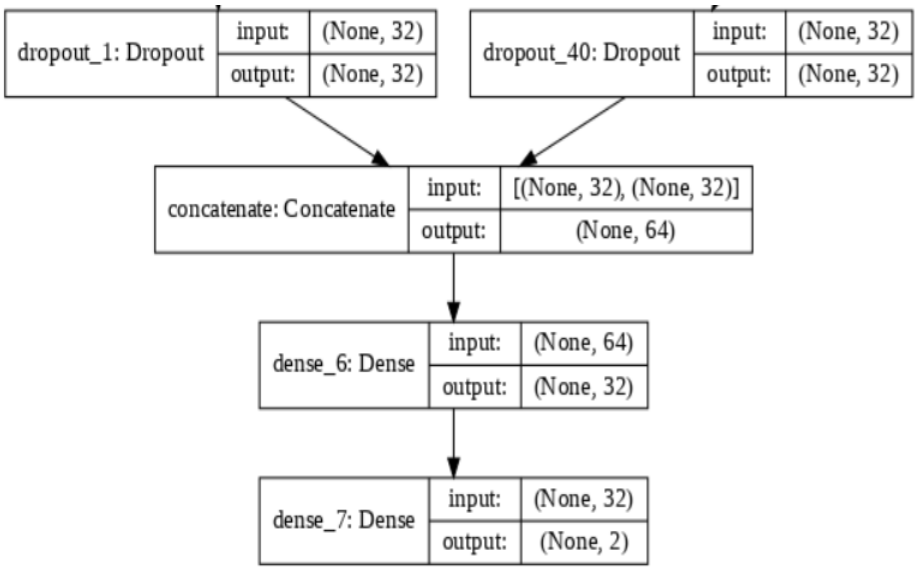


Figure 7.6: Multimodal (text + image) Model Snippet with concatenate fusion method.

The classification scores are presented in Table 7.2. The loss variations and accuracy over the epochs for training and validation sets are shown in Figure 7.8

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
8	0.9444	0.1386	0.9167	0.2176	0.9188

Table 7.2: Multimodal Model with Concatenate fusion method: Classification scores.

The confusion matrix and classification report are presented in Figure 7.7.

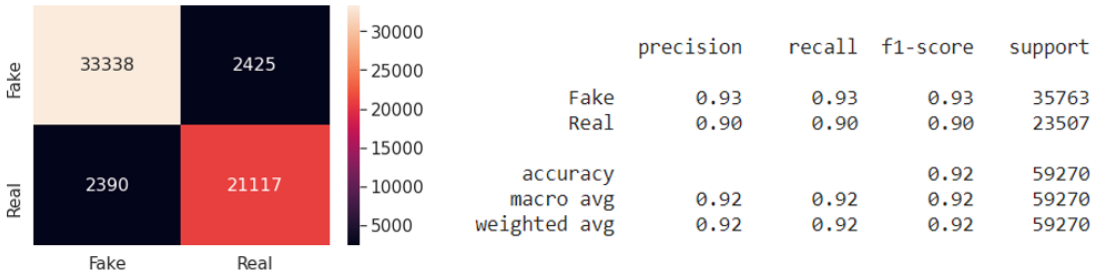


Figure 7.7: Multimodal Model with concatenate fusion method: Confusion matrix and classification report

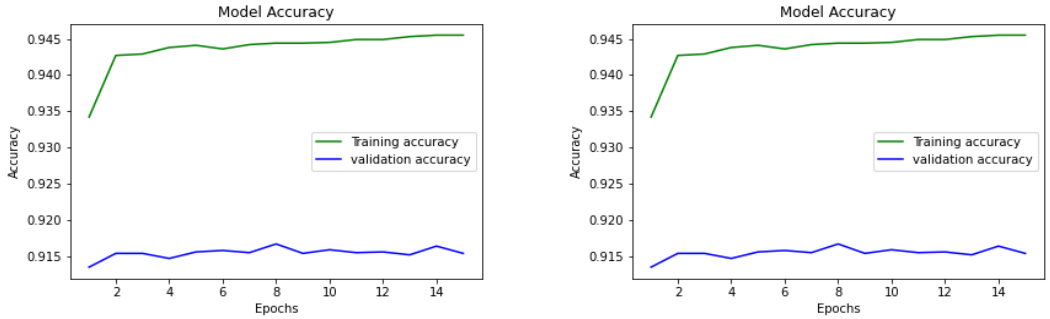


Figure 7.8: Multimodal Model with concatenate fusion method: Training and Validation graphs.

7.2 Inferences

It is observed that multimodal models outperforms unimodal models in terms of accuracy, precision, recall and F1-score. These results further validate that multi-modality helps in learning of better distinguishing features between fake and real news.

CHAPTER 8

Fake News Detection based on Image Sentiment

Fake images online tend to have strong visual impacts and often induce negative sentiments. Both tampered and the misleading images too can convey such sentiments. [Tampered fake images are those which contain manipulations. Misleading fake images do not contain such alterations but are used in false contexts intentionally.] In chapter 6, Error Level Analysis (ELA) was performed on all images in the dataset before using them for fine-tuning. However, it must be noted that ELA analysis is not tailored to bring out cues to identify misleading images (which are also a type of fake images) since these type of images do not contain tampering or alterations.

Thus, analysing the polarity of the image could be helpful. Misleading images often reflect strong negative emotions. An image which has a high negative polarity is most likely to be fake. Chapter 2 discusses key literature in the area of image polarity detection. Such analysis is popular with text, termed as sentiment analysis. In this project, sentiment analysis for images shall be carried out

In order to learn features to analyse the sentiment of an image, transfer learning on a CNN architecture (similar procedure followed for ELA images) is a suitable approach. This sections discuss details of the dataset used and the results of the transfer learning process.

8.1 Dataset for Visual Sentiment Analysis

CrowdFlower added this data set on March 27, 2015, and it contains over fifteen thousand sentiment-scored images. Images ranging from celebrity portraits to landscapes to stock photography are scored based on their typical positive/negative sentiment. Each

image was assigned one of the scores from following: positive,highly positive, negative, highly negative or neutral, . We used a subset of 1000 positive images (highly positive and positive) and 1000 negative images from this dataset (negative and highly negative). There are in total 225 validation images and 450 test images.

Some example images from dataset:

Figure 8.1: Examples of negative images in training set

Figure 8.2: Examples of positive images in training set

	Training	Validation	Test
Negative Images	1000	112	250
Positive Images	1000	113	250
Total	2000	225	450

8.2 Experiment 1 : Transfer learning with VGG19

The VGG19 model is a variation of the VGG model that has 19 layers in total (16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer).

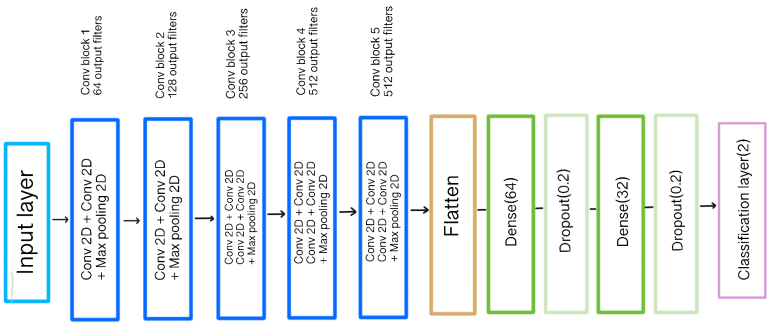


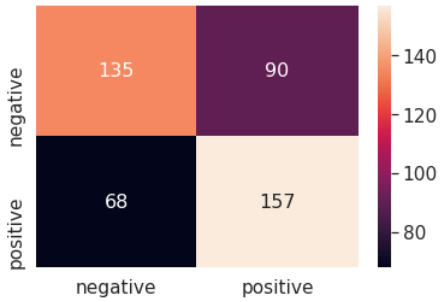
Figure 8.3: Vgg19 model architechture

The following steps were followed to fine-tune the VGG19 network :

- The Base model is instantiated and pre-trained weights from ImageNet are loaded.
- a flatten layer followed by two blocks of dense layer + Dropout layers is added on top.
- At last a softmax-activated output dense layer with two classes is added.

8.2.1 Phase 1

- All layers are frozen for base model with preladed imagenet weights.
- Adam optimizer is used, learning rate is 0.00001 and loss function is binary crossentropy.
- Batch size is set to be100 for training, 25 for validation and the model is run for 40 epochs.



Accuracy:0.6489
Precision:0.6356
Recall:0.6978
F1_Score:0.6653

	precision	recall	f1-score	support
0	0.67	0.60	0.63	225
1	0.64	0.70	0.67	225
accuracy			0.65	450
macro avg	0.65	0.65	0.65	450
weighted avg	0.65	0.65	0.65	450

Figure 8.4: VGG19 Phase 1: Confusion matrix (0-Negetive, 1-Positive) and classification report

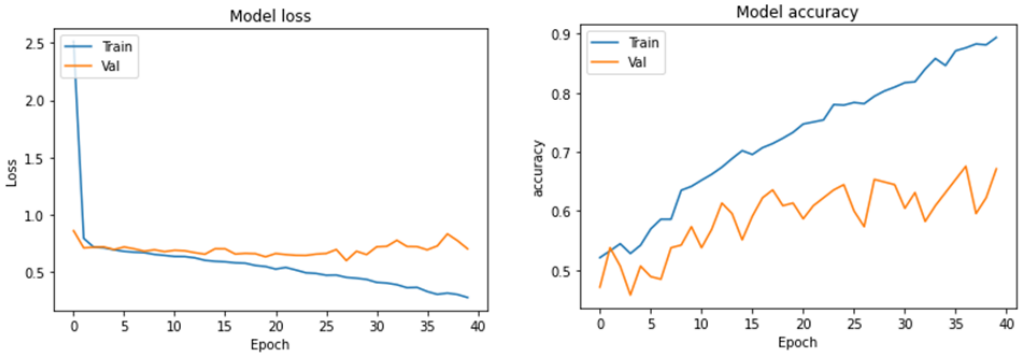


Figure 8.5: VGG19 Phase 1: Training and Validation graphs

Results

The best validation loss obtained was 0.7285 obtained in the 37th epoch and best Validation accuracy was 0.6756.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.4. It is observed that the recall for negative images is low and we can try by increasing the number of trainable parameters, the accuracy might increase.

8.2.2 Phase 2

- starting 18 layers are frozen for base model with preladed imagenet weights and trainable parameters increased up to 7,605,922.
- Adam optimizer is used, learning rate is 0.00001 and loss function is binary crossentropy.
- Batch size is set to 100 for training, 25 for validation and the model is run again for 40 epochs.

Results

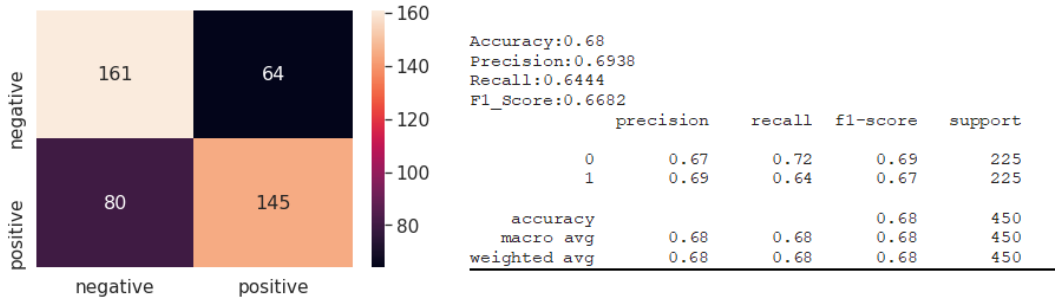


Figure 8.6: VGG19 Phase 2: Confusion matrix (0-Negetive, 1-Positive) and classification report

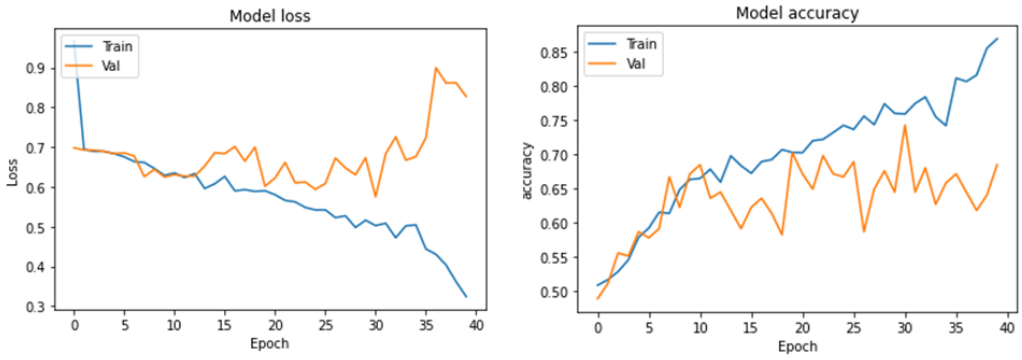


Figure 8.7: VGG19 Phase 2: Training and Validation graphs

The best validation loss obtained was 0.5751 obtained in the 31st epoch and best Validation accuracy was 0.7422. The model is evaluated against the test dataset using these weights.. The confusion matrix and classification scores are presented in Figure8.6. It is observed that the precision for positive images increased, recall for negative images increased and both test and validation accuracy increased.

8.3 Experiment 2 : Transfer learning with Resnet50

ResNet50 is a ResNet variant with 48 Convolution layers, 1 MaxPool layer, and 1 Average Pool layer.

The following steps were followed to fine-tune the Resnet50 network :

- The Base model is instantiated and pre-trained weights from ImageNet are loaded.

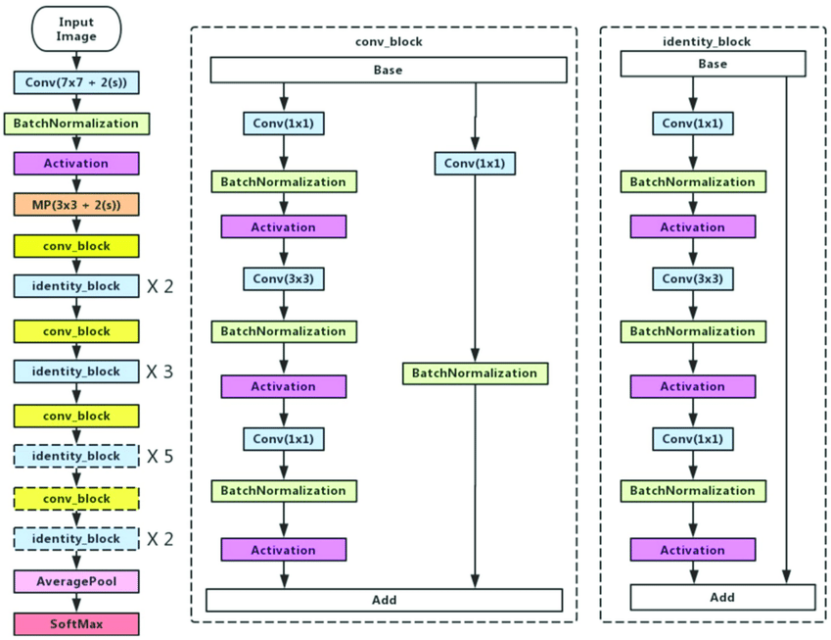


Figure 8.8: Resnet50 model architechture [7]

- a flatten layer followed by two blocks of dense layer + Dropout layers is added on top.
- Finally we add an output dense layer having two classes with softmax activation.

8.3.1 Phase 1

- Starting 170 layers are frozen for base model with preladed imagenet weights.
- Optimizer is Adam with learning rate of 0.00001 and loss function is binary crossentropy.
- Batch size is100 for training, 25 for validation and the model is run for 30 epochs.

Results

The best validation loss obtained was 0.6760 obtained in the 15th epoch and best Validation accuracy was 0.6844.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.9. It is observed that the recall for negative images is low and we can try by increasing the number of trainable parameters, the accuracy might increase.

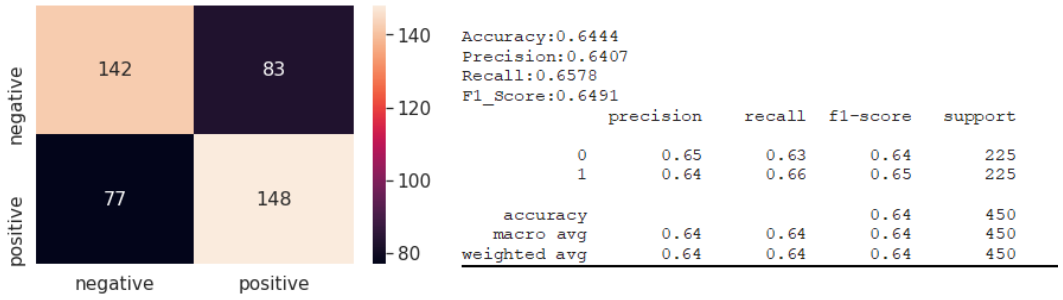


Figure 8.9: Resnet50 Phase 1: Confusion matrix (0-Negetive, 1-Positive) and classification report

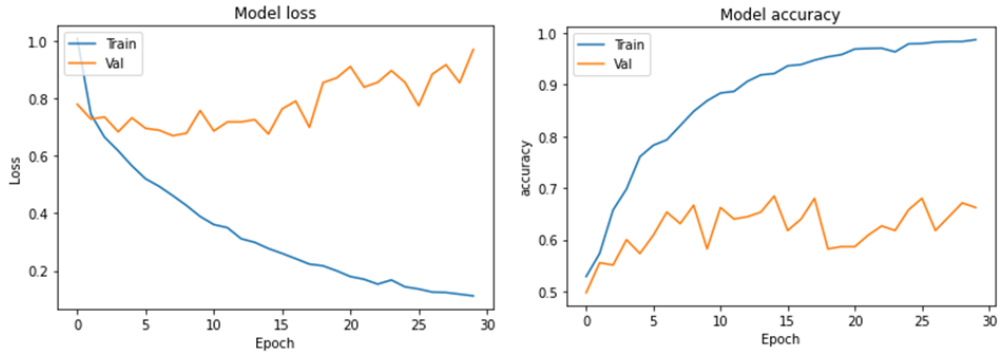


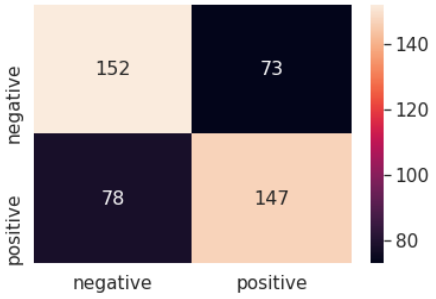
Figure 8.10: Resnet50 Phase 1: Training and Validation graphs

8.3.2 Phase 2

- starting 165 layers are frozen for base model with preladed imagenet weights and trainable parameters increased up to 7,744,674s.
- Adam optimizer is used, learning rate is 0.00001 and loss function is binary crossentropy.
- Batch size being 100 for training, 25 for validation and the model is run again for 30 epochs.

Results

The best validation loss obtained was 0.6760 obtained in the 5th epoch and best Validation accuracy was 0.7111.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.11. It is observed that the precision for positive images increased, recall for negative images increased and both test and validation accuracy increased to 71.11%.




```
Accuracy:0.6644
Precision:0.6682
Recall:0.6533
F1_Score:0.6607
              precision    recall  f1-score   support

           0       0.66      0.68      0.67       225
           1       0.67      0.65      0.66       225

    accuracy                           0.66       450
   macro avg       0.66      0.66      0.66       450
weighted avg       0.66      0.66      0.66       450
```

Figure 8.11: Resnet50 Phase 2: Confusion matrix (0-Negetive, 1-Positive) and classification report

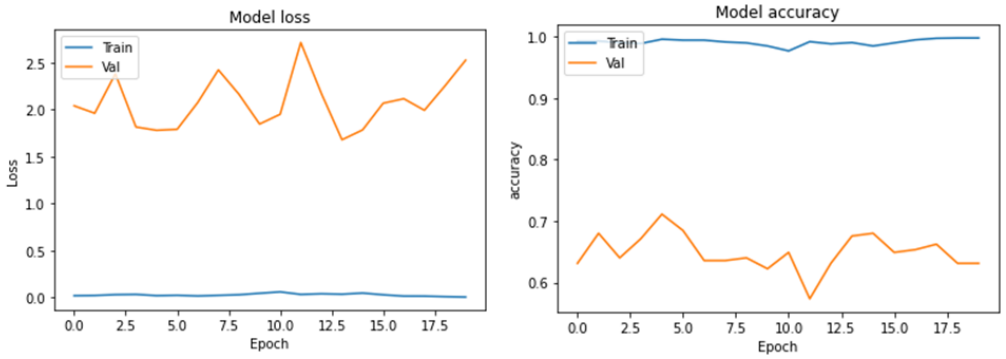


Figure 8.12: Resnet50 Phase 2: Training and Validation graphs

8.4 Experiment 3 : Transfer learning with Resnet50V2

ResNet50V2 is an updated version of ResNet50 that performs better on the ImageNet dataset than ResNet50 and ResNet101. A change was made to the propagation formulation of the connections between blocks in this model.

The following steps were followed to fine-tune the Resnet50V2 network :

- The Base model is instantiated with input shape (150, 150, 3) and pre-trained weights from ImageNet are loaded.
- a flatten layer is added followed by two blocks of dense layer + Dropout layers on top.
- At last a softmax-activated output dense layer with two classes is added

8.4.1 Phase 1

- Starting 185 layers are frozen for base model with preloaded imagenet weights.
- Optimizer is Adam, with learning rate of 0.00001.

- Batch size is set to be100 for training, 25 for validation and the model is run for 30 epochs.

Results

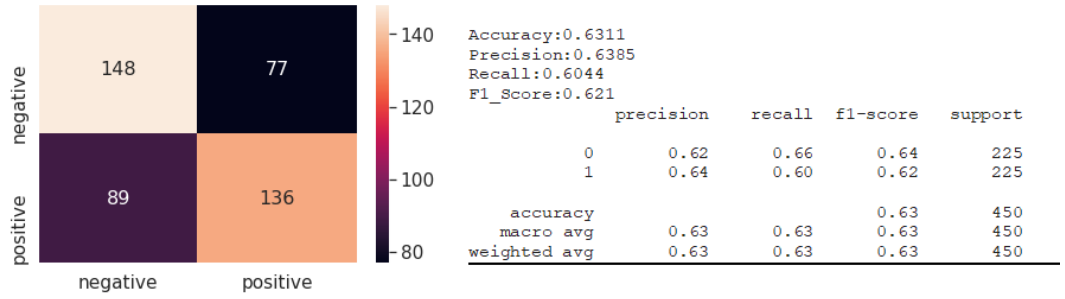


Figure 8.13: Resnet50V2 Phase 1: Confusion matrix (0-Negetive, 1-Positive) and classification report

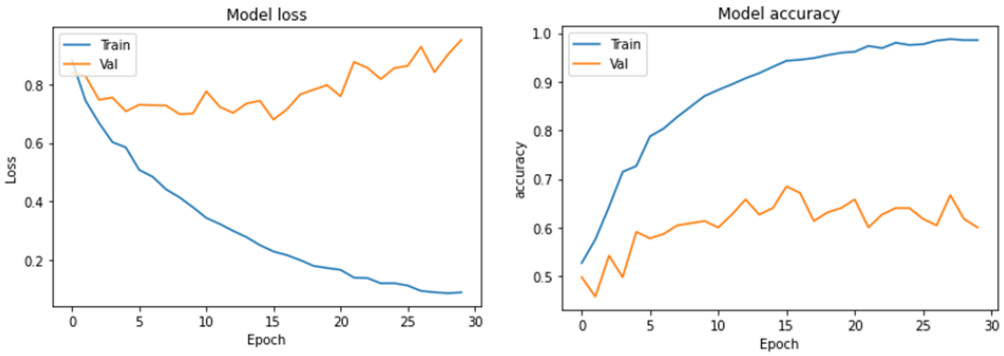


Figure 8.14: Resnet50V2 Phase 1: Training and Validation graphs

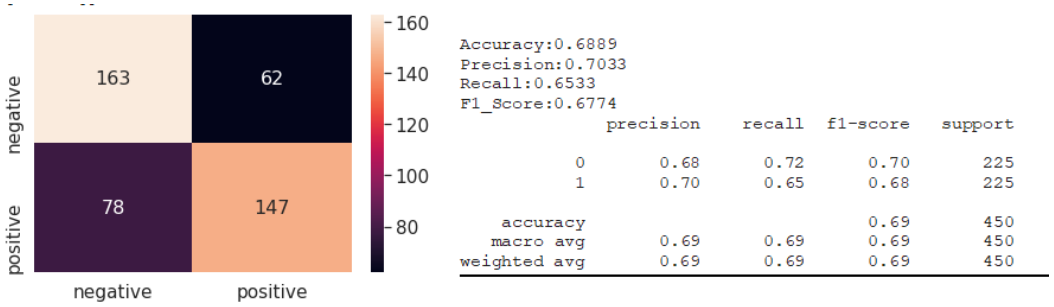
The best validation loss obtained was 0.6796 obtained in the 16th epoch and best Validation accuracy was 0.6844.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.13. It is observed that the overall accuracy is less compared to Resnet50 and we can try by increasing the number of trainable parameters, the accuracy might increase,from graph we see that accuracy might decrease with more epochs.

8.4.2 Phase 2

- starting 180 layers are frozen for base model with preladed imagenet weights and trainable parameters increased up to 6,695,074.
- Adam optimizer is used, learning rate is 0.00001 and loss function is binary crossentropy.

- Batch size is100 for training, 25 for validation and the model is run again for 30 epochs.

Results



	precision	recall	f1-score	support
0	0.68	0.72	0.70	225
1	0.70	0.65	0.68	225
accuracy			0.69	450
macro avg	0.69	0.69	0.69	450
weighted avg	0.69	0.69	0.69	450

Figure 8.15: Resnet50V2 Phase 2: Confusion matrix (0-Negetive, 1-Positive) and classification report

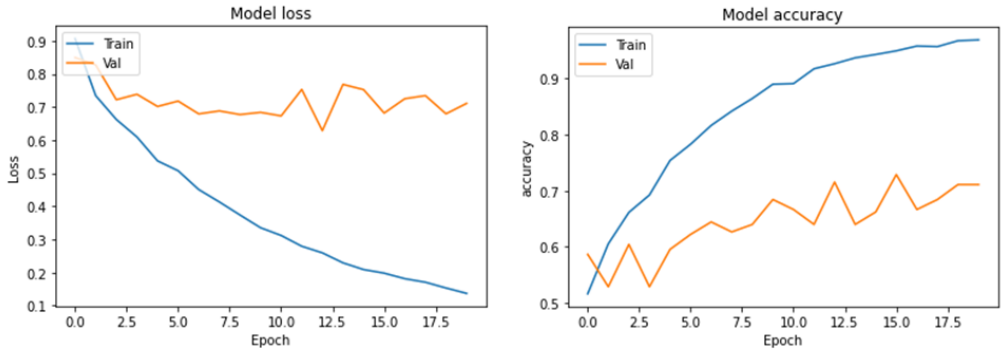


Figure 8.16: Resnet50V2 Phase 2: Training and Validation graphs

The best validation loss obtained was 0.7142 obtained in the 16th epoch and best Validation accuracy was 0.7289.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.15. It is observed that both test accuracy and validation accuracy increased, but model is starting to overfit and still this accuracy is less than what we achieved in Resnet50 phase2.

8.5 Experiment 4 : Transfer learning with InceptionV3

Inception-v3 is a 48-layer deep convolutional neural network. It's a convolutional neural network architecture from the Inception family that uses Label Smoothing, Factor-

ized 7 x 7 convolutions, and an auxiliary classifer to propagate label information lower down the network, among other improvements.

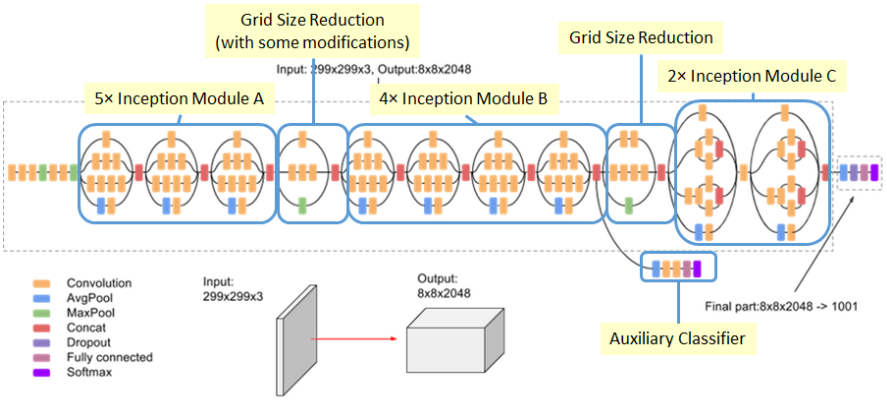


Figure 8.17: InceptionV3 model architechture
[38]

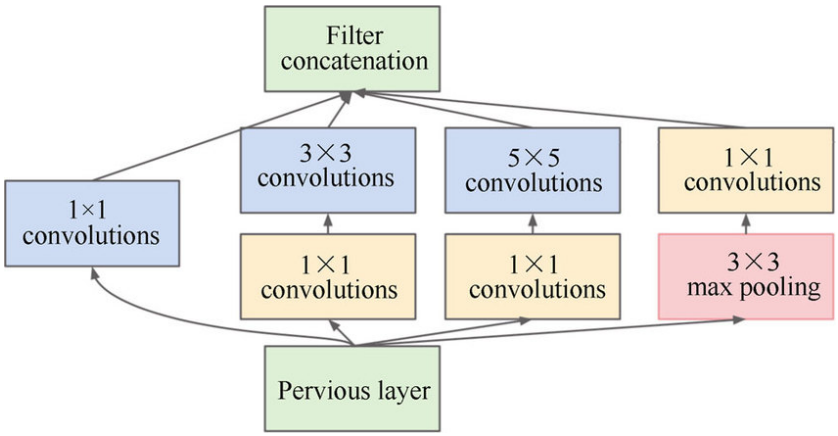


Figure 8.18: InceptionV3 module
[39]

The following steps were followed to fine-tune InceptionV3 network :

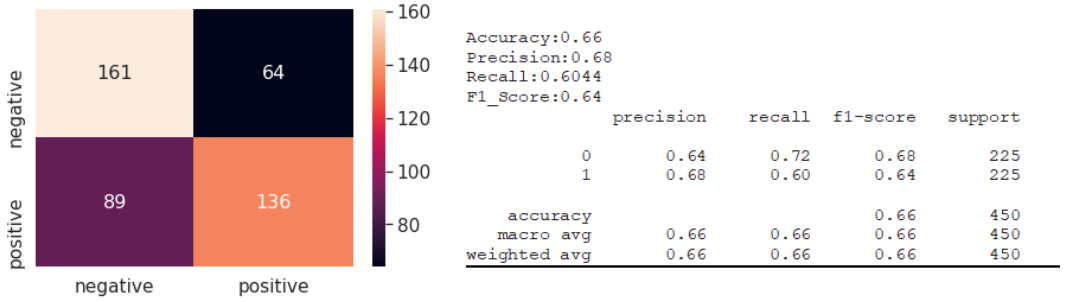
- The Base model is instantiated with input shape (150, 150, 3) and pre-trained weights from ImageNet are loaded.
- a flatten layer is added followed by two blocks of dense layer + Dropout layers on top.
- At last a softmax-activated output dense layer with two classes is added

8.5.1 Phase 1

- Starting 290 layers are frozen for base model with preloaded imagenet weights.

- Adam optimizer is used, learning rate is 0.00001 and loss function is binary crossentropy.
- Batch size:100 for training, 25 for validation and the model is run for 30 epochs.

Results



	precision	recall	f1-score	support
0	0.64	0.72	0.68	225
1	0.68	0.60	0.64	225
accuracy			0.66	450
macro avg	0.66	0.66	0.66	450
weighted avg	0.66	0.66	0.66	450

Figure 8.19: InceptionV3 Phase 1: Confusion matrix (0-Negetive, 1-Positive) and classification report

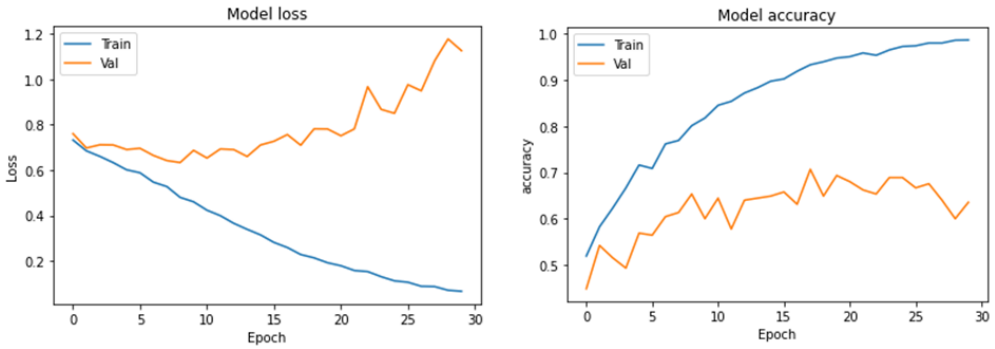


Figure 8.20: InceptionV3 Phase 1: Training and Validation graphs

The best validation loss obtained was 0.7095 obtained in the 16th epoch and best Validation accuracy was 0.7067. The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.19. It is observed We observe that for this model Recall is good and the validation accuracy has not reached saturation. Therefore Phase 2 is implemented where the model is trained with more layres unfreezed to increase the learning and no of epoch increased.

8.5.2 Phase 2

- starting 250 layers are frozen for base model with preladed imagenet weights and trainable parameters increased up to 11,723,298.

- learning rate of 0.00001 and Adam optimizer is used with a and loss function is binary crossentropy.
- Batch size: 100 for training, 25 for validation and the model is run again for 40 epochs this time.

Results

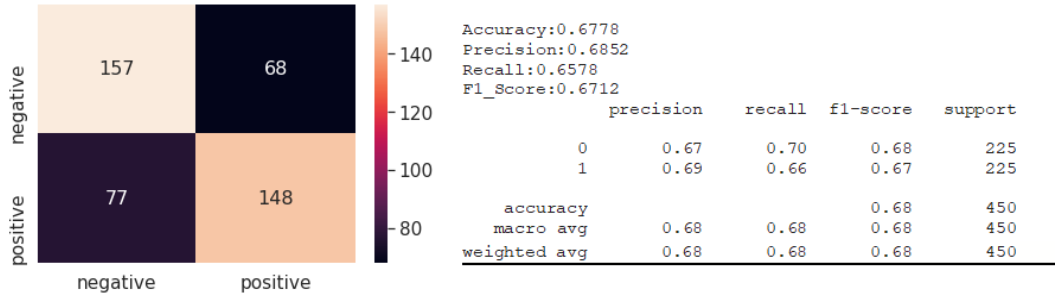


Figure 8.21: InceptionV3 Phase 2: Confusion matrix (0-Negetive, 1-Positive) and classification report

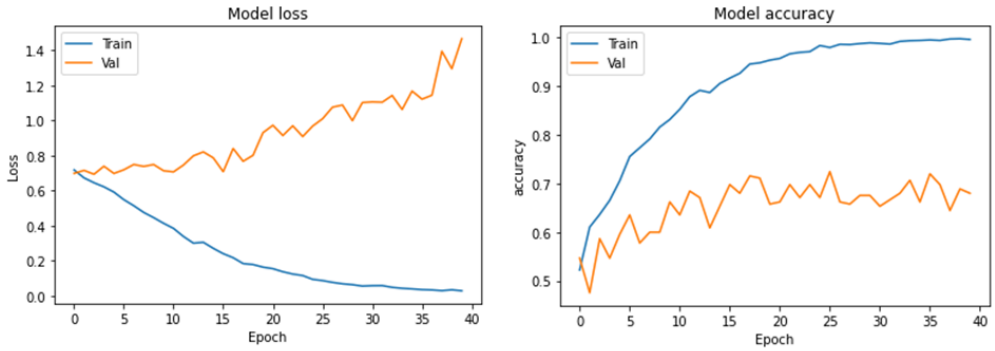


Figure 8.22: InceptionV3 Phase 2: Training and Validation graphs

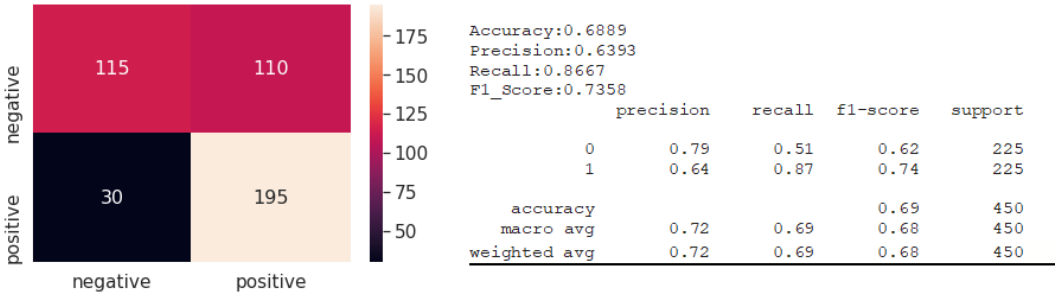
The best validation loss obtained was 1.0103 obtained in the 26th epoch and best Validation accuracy was 0.7244.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.22. It is observed that both test accuracy and validation accuracy increased,but recall and precision are similar to phase1 , so we will once try to freeze all the layers to avoid saturation in the model accuracy.

8.5.3 Phase 3

- All layers are frozen for base model with preladed imagenet weights.

- learning rate is incresed to 0.0001 to avoid overfitting and Adam optimizer is used with loss function binary crossentropy.
- Batch size is 100 for training, 25 for validation and the model is run for 30 epochs.

Results



	precision	recall	f1-score	support
0	0.79	0.51	0.62	225
1	0.64	0.87	0.74	225
accuracy			0.69	450
macro avg	0.72	0.69	0.68	450
weighted avg	0.72	0.69	0.68	450

Figure 8.23: InceptionV3 Phase 3: Confusion matrix (0-Negetive, 1-Positive) and classification report

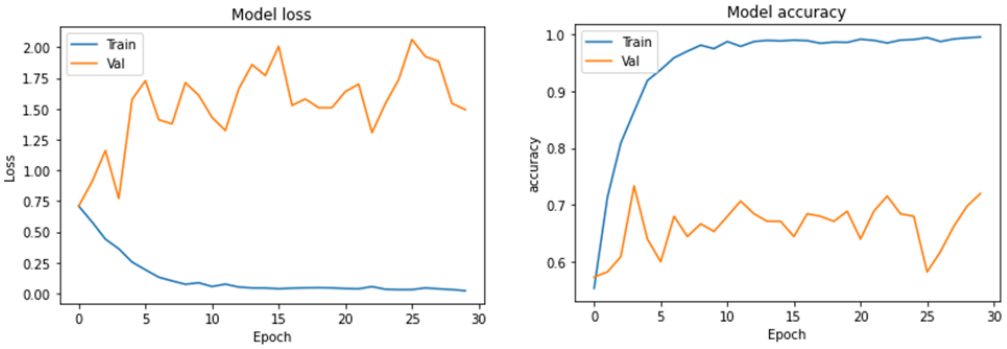


Figure 8.24: InceptionV3 Phase 3: Training and Validation graphs

The best validation loss obtained was 0.7711 obtained in the 3rd epoch and best Validation accuracy was 0.7333. The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.24. It is observed that overall accuracy has incresed, but recall for negative images went down, also even with a high learning rate and less epochs, model accuracy is reacching saturation.

8.6 Experiment 5 : Transfer learning with Xception

Xception is a Depthwise Separable Convolutions-based deep convolutional neural network design. It was created by Google's research team. A depthwise separable convolution can be thought of as an Inception module with the most towers possible. In most traditional classification problems, the Xception architecture outperformed VGG-16, ResNet, and Inception V3.

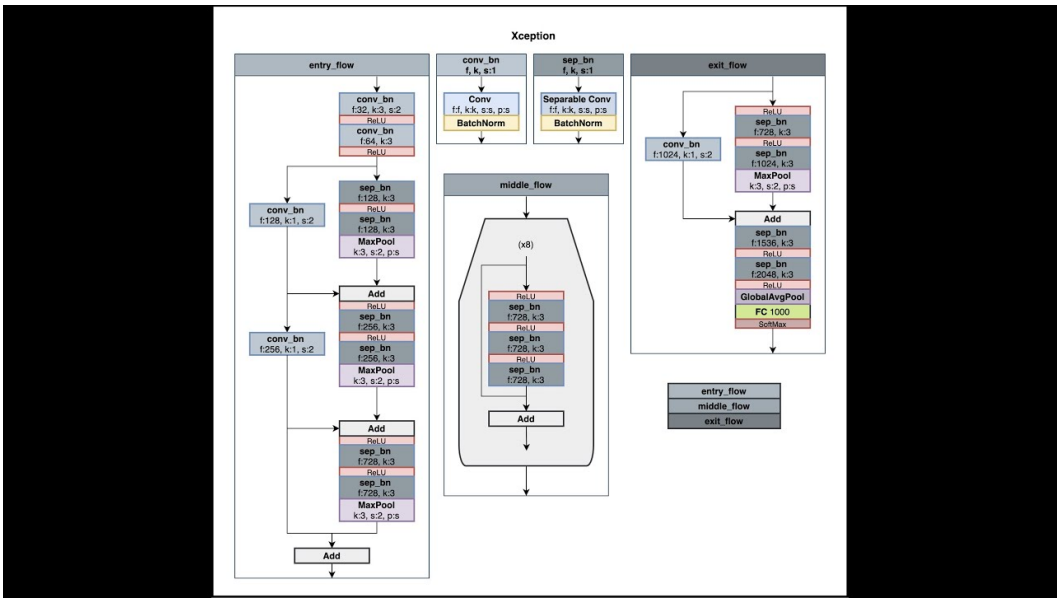


Figure 8.25: Xception model architechture
[40]

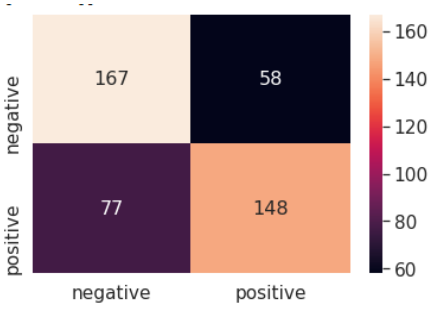
The following steps were followed to fine-tune Xception network :

- The Base model is instantiated with input shape (150, 150, 3) and pre-trained weights from ImageNet are loaded.
- a flatten layer is added followed by two blocks of dense layer + Dropout layers on top.
- At last a softmax-activated output dense layer with two classes is added

8.6.1 Phase 1

- All layers are frozen for base model with preloaded imagenet weights.
- Optimizer is Adam, learning rate is 0.0001.
- Batch size is set as 100 for training, 25 for validation and the model is run for 30 epochs.

Results



Accuracy:0.7
Precision:0.7184
Recall:0.6578
F1_Score:0.6868

	precision	recall	f1-score	support
0	0.68	0.74	0.71	225
1	0.72	0.66	0.69	225
accuracy			0.70	450
macro avg	0.70	0.70	0.70	450
weighted avg	0.70	0.70	0.70	450

Figure 8.26: Xception Phase 1: Confusion matrix (0-Negetive, 1-Positive) and classification report

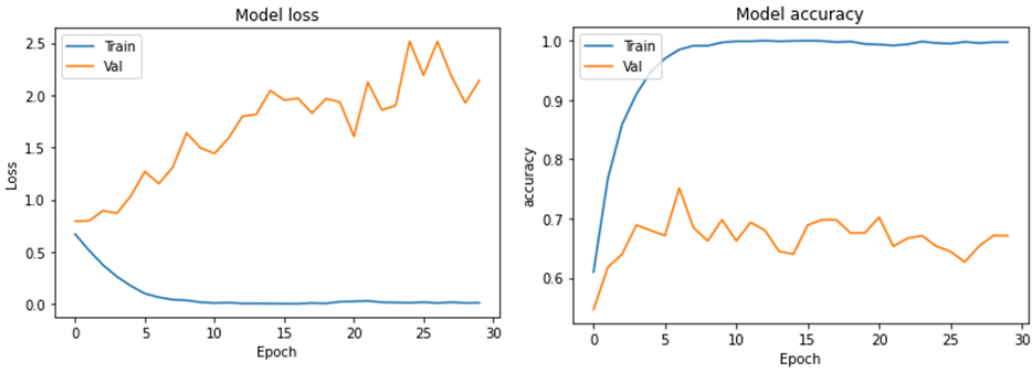


Figure 8.27: Xception Phase 1: Training and Validation graphs

The best validation loss obtained was 1.1537 obtained in the 7th epoch and best Validation accuracy was 0.7511. The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.26. It is observed We observe that for this model model accuracy is best till yet of all models,we will try to increase no of epochs to see if accuracy increases. Recall value for negative images is increased.

8.6.2 Phase 2

- starting 115 layers are frozen for base model with preladed imagenet weights and trainable parameters increased up to 10,067,394.
- learning rate of 0.0001 and Adam optimizer is used with a and binary crossentropy loss function.
- Batch size is set 100 for training, 25 for validation and the model is run again for 40 epochs this time.

Results

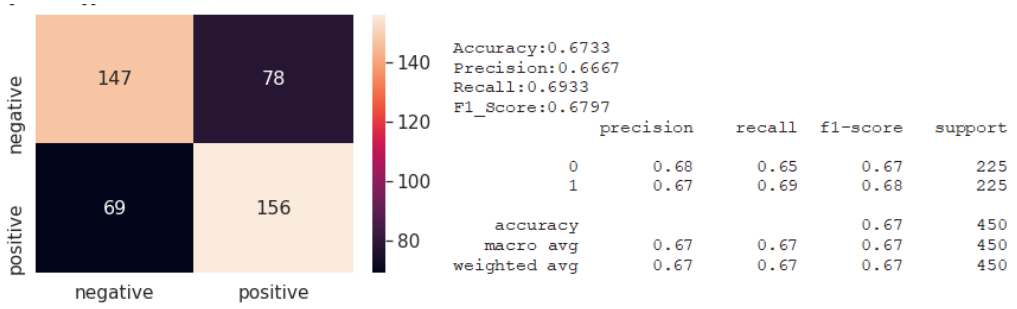


Figure 8.28: Xception Phase 2: Confusion matrix (0-Negetive, 1-Positive) and classification report

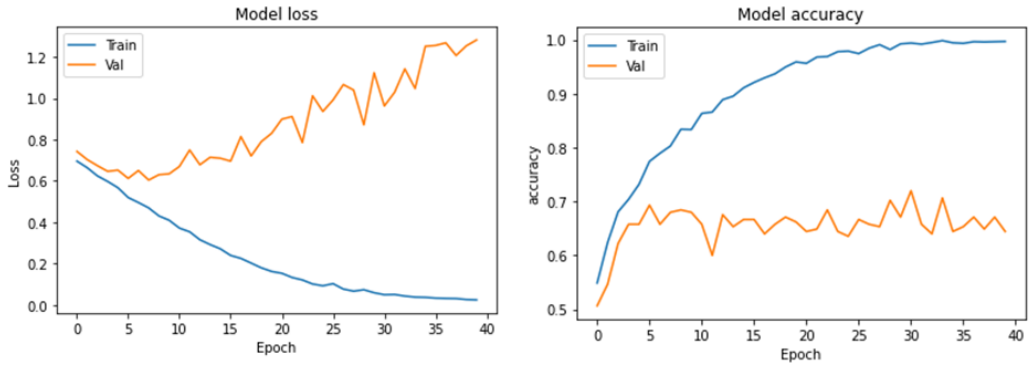


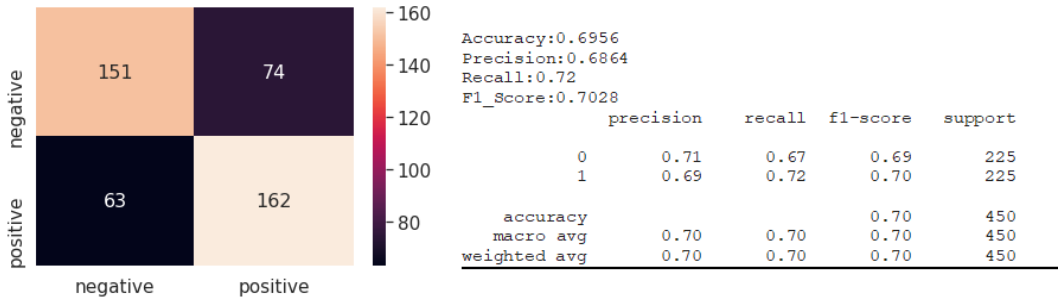
Figure 8.29: Xception Phase 2: Training and Validation graphs

The best validation loss obtained was 0.9625 obtained in the 30th epoch and best Validation accuracy was 0.7200.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.29. It is observed that both test accuracy and validation accuracy went down,recall and precision have also decreased.

8.6.3 Phase 3

- All layers are frozen for base model with preladed imagenet weights.
- after flatten layer,two modules of dense layer(266 and 128) + Dropout(0.2) layers are added
- learning rate is incresed of 0.0001 and Adam optimizer is used with loss function as binary crossentropy.
- Batch size is set 100 for training, 25 for validation and the model is run for 30 epochs.

Results



	precision	recall	f1-score	support
0	0.71	0.67	0.69	225
1	0.69	0.72	0.70	225
accuracy			0.70	450
macro avg	0.70	0.70	0.70	450
weighted avg	0.70	0.70	0.70	450

Figure 8.30: Xception Phase 3: Confusion matrix (0-Negetive, 1-Positive) and classification report

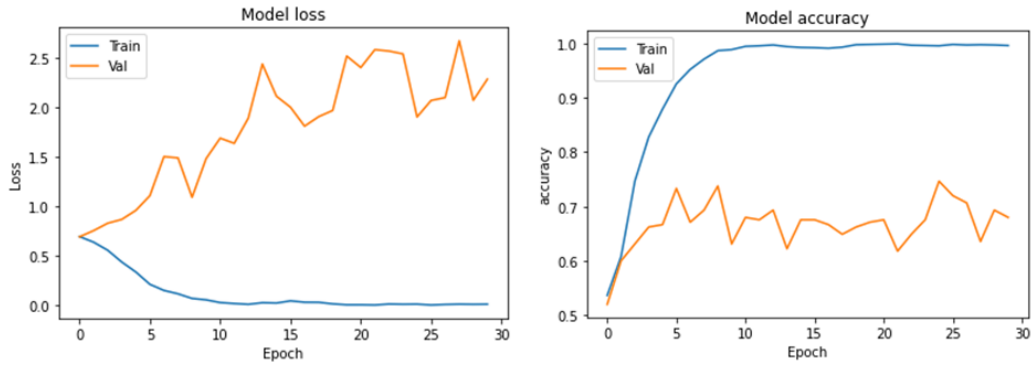


Figure 8.31: Xception Phase 3: Training and Validation graphs

The best validation loss obtained was 1.9060 obtained in the 24th epoch and best Validation accuracy was 0.7467.The model is evaluated against the test dataset using these weights. The confusion matrix and classification scores are presented in Figure8.31. It is observed that overall accuracy is comparable to phase1 but, to avoid more complexity we shall use phase1 model.

CHAPTER 9

Proposed Framework for Fake News Detection

In the previous chapter, we looked at using visual sentiment analysis to determine an image's polarity. This was identified as a potential step to improve fake image identification since majority of fake images carry strong negative visual impacts.

In this chapter, a new architecture is proposed which combines the power of visual sentiment analysis with previously identified multi-modal models in Chapter 7. We implemented and evaluated multi-modality (i.e. text and image) models that were trained on the Fakeddit dataset, which used both the textual (image captions) and the visual information (images) in posts to determine whether they were fake or real.

9.1 Proposed Architecture

We present our proposed framework for multimodal fake news detection in this section. Figure 9.1 depicts an overview of our approach.

Three function extractors make up the entire model: two Visual network Φ_{im} & Φ_{ip}, a Textual network Φ_t ; and one auxiliary classifiers: a Multimodal classifier C_m.

Visual Network for image manipulation detection: The objective is to extract a representation of the input image with Visual network Φ_{im}. The network will attempt to learn visual features for detecting image manipulation.

Visual Network for image polarity detection: The objective is to extract a representation of the input image with the Visual network Φ_{ip}. The network will attempt to learn visual features for image polarity detection.

Textual Network: The Φ_t Textual Network's goal is to is to extract a representation of the input text. The network will try to learn textual features for fake news detection.

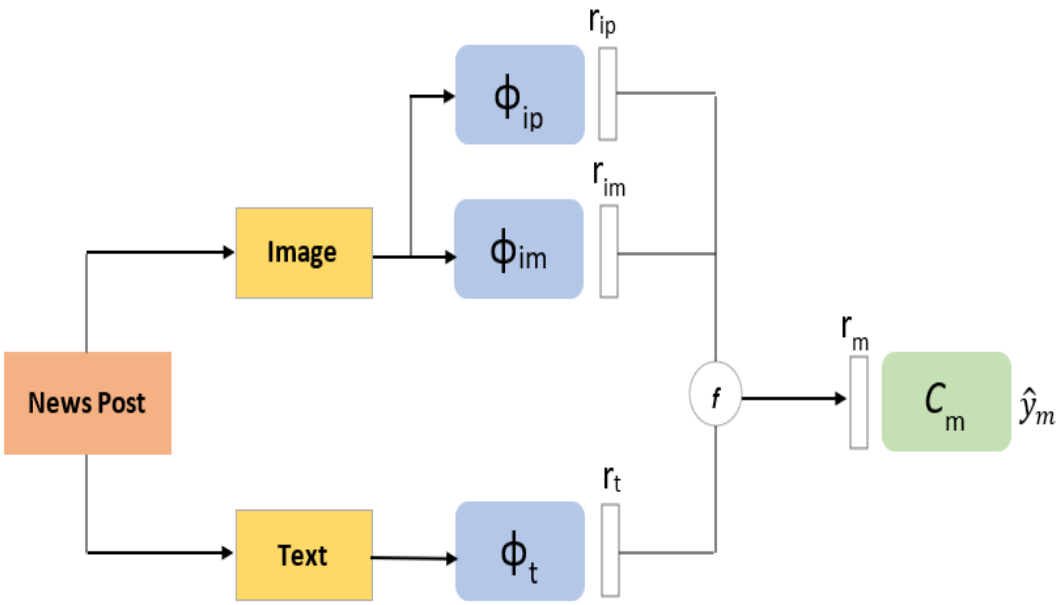


Figure 9.1: Overview of the proposed approach

Multimodal Classifier: Finally, this classifier makes use of all of the modalities' representations r_{im}, r_{ip} and r_t to predict. An initial step is to do the fusion of representations with a function f(r_{im}, r_{ip}, r_t).Several methods have been suggested to achieve this aim, ranging from basic (e.g. concatenation, summation) to more complex.

Our main goal is to identify fake news from both modalities of a given piece of news on their own, without taking into account any other sub-tasks.

9.2 Implementation Details

In Chapter 5, we fine-tuned the BERT model using just the textual information i.e., the image captions present in the post were used for fine-tuning the BERT model.

Similar in Chapter 6, the Xception model was fine-tuned using the image present in the post. In Chapter 8, the Xception model was fine-tuned for visual sentiment analysis. The model reported high precision, recall and F1-scores for both positive and negative images, thereby promising to aid in fake image identification.

To use the power of all 3 techniques, an ensemble model was designed to improve the identification of fake news - with

- Xception network to help in identifying images with high digital alterations.
- The power of language models will be incorporated into BERT, which means the model will learn contextual knowledge. Contextual information is essential in addition to content information since real-world texts, images, and videos are complex.
- and visual sentiment analysis to learn features which distinguish an image with a negative sentiment from that which induces positive emotions - thereby identify misleading and also tampered fake images with high confidence.

The design of the ensemble construction is depicted in Figure 9.2 The leftmost vertical branch consists of layers from the BERT model fine-tuned on image captions, the middle branch is composed of layers from the Xception model fine-tuned on Fakeddit dataset images, and the rightmost vertical branch is comprised of layers from the Xception model fine-tuned on Sentiment dataset images. All layers from these models except the last classification layers are included in this proposed ensemble model. Finally, the last module is a multimodal fusion module that combines representations from various modalities (such as text and image) to form a news feature vector. For fake news classification, this news representation is fed into a completely connected neural network with softmax activation.

The proposed ensemble model is loaded with the best weights obtained from 3 models which were trained independently in previous chapters - fine-tuning Xception on Fakeddit dataset images, fine-tuning BERT on image captions(i.e. text) and fine-tuning Xception network for sentiment analysis. The ensemble model is fine-tuned again on Fakeddit dataset samples that have both image and text. The following phases detail some experiments performed to obtain optimal weights.

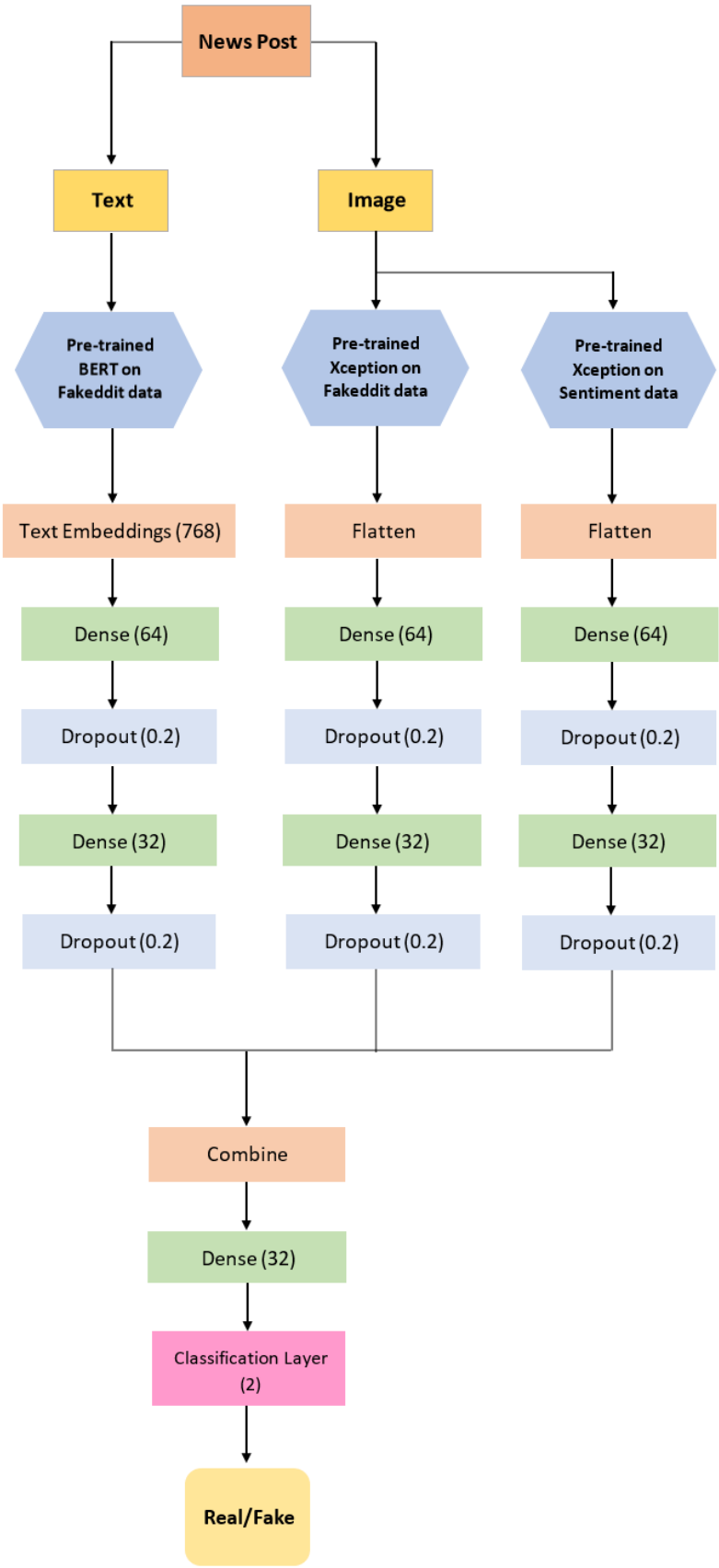


Figure 9.2: Proposed Model Design

9.2.1 Phase 1

- All layers in the Xception branch are made untrainable until the Flatten layer, and all layers in the BERT branch are made untrainable until the Text embeddings. In other words, we made the base models in both branches untrainable. We will train only the fully connected dense layers.
- The whole sentiment branch was made untrainable till the Merge/combine layer.
- Both modalities' 32-dimensional vectors are combined using **maximum fusion method** and fed into a fully connected neural network classifier with a 32-layer hidden layer and a 2-layer classification layer with softmax activation. (Refer Fig. 9.3)
- Batch size = 256
- Adam optimiser is set with 0.0005 as learning rate and model is trained for 20 epochs.

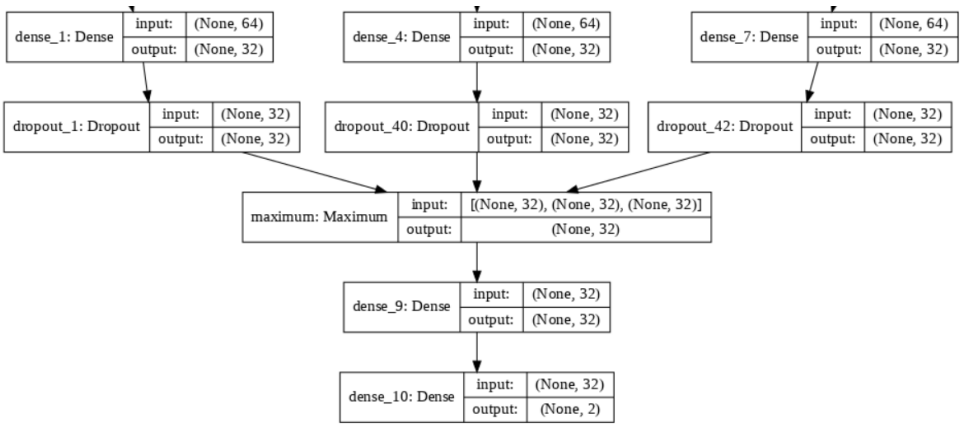


Figure 9.3: Proposed method with Max fusion method.

Results

The best validation loss was obtained, at the 12th epoch out of a total of 20 epochs. The model is evaluated against the test dataset using the weights obtained at this stage. The classification scores are presented in Table 9.1. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 9.4.

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
12	0.9443	0.1376	0.9168	0.2195	0.9194

Table 9.1: Proposed Model with Max fusion method: Classification scores.

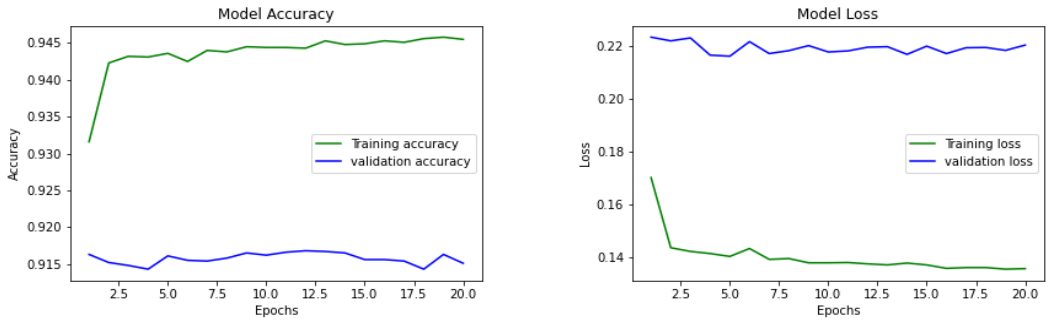
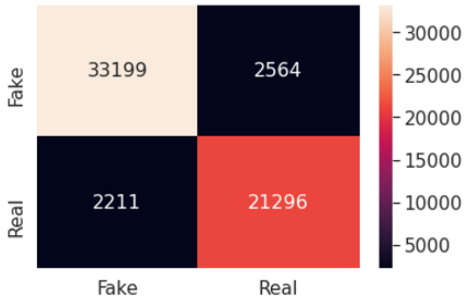


Figure 9.4: Proposed Model with Max fusion method: Training and Validation graphs.

The confusion matrix and classification report are presented in Figure 9.5. It is observed that the accuracy, precision, recall and F1-score have slightly improved as compared to the multimodal models in Chapter 7.



	precision	recall	f1-score	support
Fake	0.94	0.93	0.93	35763
Real	0.89	0.91	0.90	23507
accuracy			0.92	59270
macro avg	0.92	0.92	0.92	59270
weighted avg	0.92	0.92	0.92	59270

Figure 9.5: Proposed Model with Max fusion method: Confusion matrix and classification report

9.2.2 Phase 2

For this phase, the following changes are made:

- Both modalities' 32-dimensional vectors are combined using **concatenation fusion method** and fed into a fully connected neural network classifier with a 32-layer hidden layer and a 2-layer classification layer with softmax activation (Refer Fig. 9.6)
- No changes made in optimizer settings.

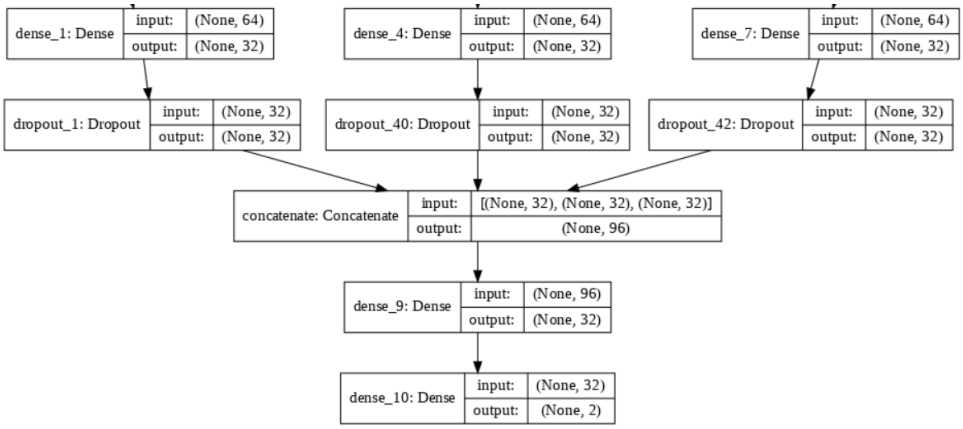


Figure 9.6: Proposed method with Concatenate fusion method.

Results

The best validation loss was obtained, at the 13th epoch out of a total of 20 epochs.The model is evaluated against the test dataset using the weights obtained at this stage. The classification scores are presented in Table 9.2. The accuracy and loss variations over the epochs for training and validation sets are shown in Figure 9.8

epoch	accuracy	loss	val_accuracy	val_loss	test_accuracy
13	0.9448	0.1366	0.9170	0.2231	0.9187

Table 9.2: Proposed Model with concatenate fusion method: Classification scores.

The confusion matrix and classification report are presented in Figure 9.7. There is slight improvement in validation accuracy and recall for fake images.

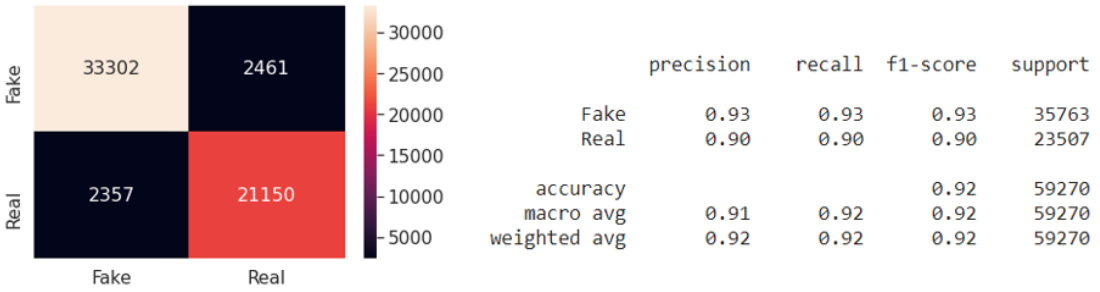


Figure 9.7: Proposed Model with concatenate fusion method: Confusion matrix and classification report

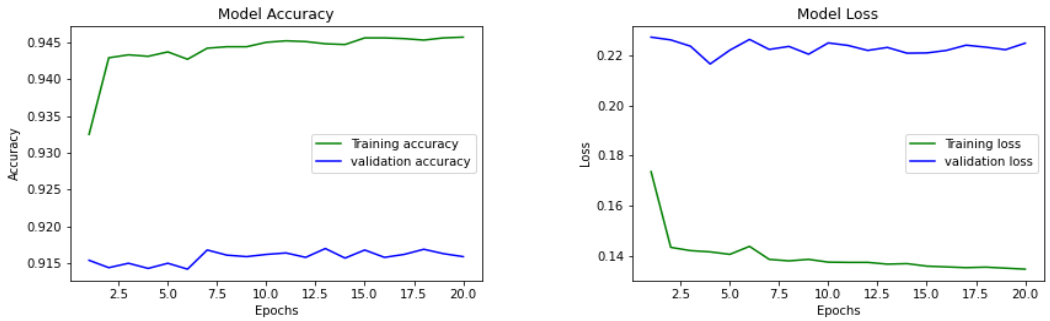


Figure 9.8: Proposed Model with concatenate fusion method: Training and Validation graphs.

9.3 Performance Comparison

Table 9.3 displays both the baselines results and our proposed approach results on Fakeddit dataset. On fake news, we report the validation and test accuracy of our model, as well as the precision, recall, and F1-score We can see how much better our proposed method performs compared to the baseline methods.

9.4 Inferences

We compare and contrast our findings with those of other methods in this section. We can observe that the **In terms of accuracy, precision, recall, and F1 score, the proposed method outperforms the current methods overall.** We can infer from the table that multimodal models outperform unimodal models. It leads to learning of better distinguishing features between fake and real news. When it comes to assessing the accuracy of news, data from various sources complements each other.

In this context of fake news images identification, a good recall score is crucial since we would not want to miss flagging a fake news image. At the same time, we also need to be reasonably precise with our predictions. **Our proposed methods have a high recall, precision, and an F1-score of $\sim$93%.**

Multi-modal models	Val accuracy	Test accuracy	Precision	Recall	F1 score
Baseline models					
InferSent + VGG16 with maximum	0.8655	0.8658	N/A	N/A	N/A
InferSent + EfficientNet with maximum	0.8328	0.8339	N/A	N/A	N/A
InferSent + ResNet50 with maximum	0.8888	0.8891	N/A	N/A	N/A
BERT + VGG16 with maximum	0.8694	0.8699	N/A	N/A	N/A
BERT + EfficientNet with maximum	0.8334	0.8318	N/A	N/A	N/A
BERT + ResNet50 with maximum	0.8929	0.8909	N/A	N/A	N/A
BERT + ResNet50 with concatenate	0.8564	0.8568	N/A	N/A	N/A
My models					
BERT + Xception with maximum	0.9161	0.9187	0.9343	0.9307	0.9325
BERT + Xception with concatenate	0.9167	0.9188	0.9331	0.9322	0.9326
Proposed Method with maximum	0.9168	**0.9194**	**0.9376**	0.9283	**0.9329**
Proposed Method with concatenate	**0.9170**	0.7987	0.9339	**0.9312**	0.9325

Table 9.3: Performance of the proposed model in comparison to the baselines

CHAPTER 10

Conclusions and Future Work

10.1 Conclusion

1. The project highlighted various issues with social media's convenience and openness. It aided the spread of fake news and continues to have disastrous consequences on society today, instilling fear, panic, and violent clashes among the public. These negative impact lay the foundation for the motivation and need for an automated fake news detection system.

2. Manually assessing the genuineness of news is difficult and time-consuming. As a result, we test our framework for fake news detection using the Fakeddit dataset, a publicly accessible dataset.

3. Fake news potentially differs from the truth in terms of writing style and quality, quantity such as word counts, and sentiments expressed. We implemented and evaluated various text modality models. The models were fine-tuned on the Fakeddit dataset and used only the textual information (image caption) in posts to determine whether they were fake or not. The fine-tuned BERT model outperforms the baseline model, with an accuracy of 89.31% versus 86.44% (baseline accuracy).

4. Given that visual content is an important promoter for fake news propaganda and provides numerous cues for detecting fake news, we implemented various image modality models. The fine-tuned Xception network outperforms the baseline model, with an accuracy of 82.32% versus 80.70% (baseline accuracy).

5. Using either text or an image alone may not be enough to detect falsification. Therefore, we implemented various multi-modality (text + image) models to address this constraint. This significantly improves the recall of fake images, with a score of 0.93 when compared to recall values obtained from the uni-modal models. The fine-tuned multi-modal network (Xception + BERT with concatenation technique) outperforms the baseline model, with an accuracy of 91.88% versus 85.68% (baseline accuracy).

6. Fake images online tend to have strong visual impacts and often induce negative sentiments. Visual sentiment analysis is a relatively new research area in which image polarity detection analyses the sentiment of a given image - to identify whether an image reflects positive or negative sentiments. Thus, in this project, the approach of visual sentiment analysis is augmented with the previously fine-tuned multimodality models.

7. The project proposed an ensemble model consists of two CNN architectures to learn visual features and one language model (BERT) to learn text features. The model utilizes the spatial properties of CNNs to look for physical alterations in an image as well as analyse if the image reflects a negative sentiment, since fake images often exhibit either one or both caracteristics. The BERT captures contextual information.

8. Unlike traditional image forensic techniques, the model has been trained to identify both tampered and untampered but misleading images. The proposed model performs better than the baseline, having an accuracy of 91.94% with an Precison, Recall and F1 score of 93%.

10.2 Future Scope

1. One way to experiment and improve the performance is by incorporating user and social context-based features. It is encouraged to predict false news from many perspectives together, so that their strengths can be combined.

2. In future research, the use of the metadata and comment data provided can help track the credibility of a user.

3. Cross-domain Generalization: Improve the model's prediction performance across domains, topics, websites, and languages.

4. Explainable Fake News Detection: Attention mechanisms can be explored to ensure that different features are appropriately highlighted. Techniques such as feature map visualisations can be explored to improve interpretability of the model.

5. Nowadays, fake news is propagated via memes (text embedded in an image). One can develop a model that can predict the 'fakeness'/genuineness of such images

REFERENCES

[1] C. Silverman, "This analysis shows how viral fake election news stories outperformed real news on facebook," 2016. [Online]. Available: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

[2] L. Clever, D. Assenmacher, K. Müller, M. V. Seiler, D. M. Riehle, M. Preuss, and C. Grimme, "Fakeyou! – a gamified approach for building and evaluating resilience against fake news," *arXiv preprint arXiv:2003.07595*, 2020.

[3] R. Thakur and R. Rohilla, "Recent advances in digital image manipulation detection techniques: A brief review," *Forensic Science International*, vol. 312, p. 110311, 2020.

[4] Z. Elhamraoui, "Inceptionresnetv2 simple introduction," 2020. [Online]. Available: https://medium.com/@zahraelhamraoui1997/inceptionresnetv2-simple-introduction-9a2000edcdb6

[5] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017.

[6] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. K. et al., "Verifying multimedia use at mediaeval 2015," vol. 3, no. 3, p. 7, 2015.

[7] Q. Ji, J. Huang, W. He, and Y. Sun, "Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images," *Algorithms*, vol. 12, p. 51, 02 2019.

[8] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[10] K. H. Jamieson and J. N. Cappella, "Echo chamber: Rush limbaugh and the conservative media establishment," *Oxford University Press*, 2008.

[11] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 175–220, 1998.

[12] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–34, 2020.

[13] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," *Disinformation, Misinformation, and Fake News in Social Media*, pp. 141–161, 2020.

[14] A. Chowdhury, "Fake news in the time of coronavirus: A boom study," 2020. [Online]. Available: https://www.boomlive.in/fact-file/fake-news-in-the-time-of-coronavirus-a-boom-study-8008

[15] P. V. Shah, "Multimodal fake news detection using a cultural algorithm with situational and normative knowledge," M.Sc. Thesis, University of Windsor, July 2020, https://scholar.uwindsor.ca/etd/8396.

[16] Z. Jin, J. Cao, J. Luo, and Y. Zhang, "Image credibility analysis with effective domain transferred deep networks," *arXiv preprint arXiv:1611.05328*, 2016.

[17] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 518–527, 2019.

[18] F. Lago, Q.-T. Phan, and G. Boato, "Visual and textual analysis for image trustworthiness assessment within online news," *Security and Communication Networks*, vol. 2019, 2019.

[19] D. Cozzolino and L. Verdoliva, "Camera-based image forgery localization using convolutional neural networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1372–1376, 2018.

[20] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 101–117, 2018.

[21] Y. WuEmail, W. Abd-Almageed, and P. Natarajan, "Busternet: Detecting copy-move image forgery with source/target localization," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 168–184, 2018.

[22] D. Cozzolino and G. P. andLuisa Verdoliva, "Splicebuster: A new blind image splicing detector," *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2015.

[23] A. Ortis, G. M. Farinella, and S. Battiato, "A survey on visual sentiment analysis," *IET Image Process*, vol. 14, pp. 1440–1456, 2020.

[24] G. W. Allport and L. Postman, "The psychology of rumor," 1947.

[25] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," *Proceedings of the 20th international conference on World wide web*, pp. 675–684, 2011.

[26] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," *2015 IEEE 31st International Conference on Data Engineering*, pp. 651–662, 2015.

[27] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3818–3824, 2016.

[28] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 312–320, 2019.

[29] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," *The World Wide Web Conference*, pp. 2915–2921, 2019.

[30] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816, 2017.

[31] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 849–857, 2018.

[32] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pp. 5–10, 2016.

[33] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, vol. 6, pp. 417–422, 2006.

[34] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," *Proceedings of the 18th ACM international conference on Multimedia*, pp. 715–718, 2010.

[35] T. Chen, D. Borth, T. Darrell, and S. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[36] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 381–388, 2015.

[37] N. Krawetz, "A picture ' s worth . . . digital image analysis and forensics version 2," *Hacker Factor Solutions*, vol. 6, no. 2, p. 2, 2007.

[38] DeepAI, "Inception module definition," 2020. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/inception-module

[39] B. Zhao, "Inception module of googlenet." 2021. [Online]. Available: https://www.researchgate.net/profile/Bo-Zhao-67/publication/312515254/figure/fig3/AS:489373281067012@1493687090916/nception-module-of-GoogLeNet-This-figure-is-from-the-original-paper-10.png

[40] M. A. Intelligence, “Cnn architectures - xception implementation | mlt,” 2020. [Online]. Available: https://i.ytimg.com/vi/nMBCSroJ7bY/maxresdefault.jpg