Master Thesis

Multi-Label, Multi-Task Deep Learning Approach Towards Detection The Differences Between Real And Fake Emotions

Dimitrios Gagatsis

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Data Science for Decision Making
at the Department of Advanced Computing Sciences
of the Maastricht University

Thesis Committee:

Mirela C. Popa
Alexia Briassouli

Maastricht University
Faculty of Science and Engineering
Department of Advanced Computing Sciences

March 26, 2023

Contents

1 Introduction **2**

2 Literature Review **5**

3 Methodology **10**
3.1 Single-Task Learning Approach 10
3.2 Multi-Task Learning Approach 13
3.2.1 Combined Accuracy 14
3.2.2 Loss Function 15
3.2.3 Combined Loss 16
3.3 Temporal Learning Approach 16

4 Results **19**
4.1 Data 19
4.1.1 Data Augmentation 20
4.2 Evaluation Metrics 20
4.2.1 Confusion Matrix 20
4.2.2 Accuracy 21
4.2.3 Precision 21
4.2.4 Learning Rate Scheduler 21
4.3 Single-Task approach results 22
4.4 Multi-Task approach results 23
4.5 Temporal Learning approach results 26

5 Discussion **27**

6 Conclusion **29**

Abstract

The ability to accurately recognize genuine emotions has important implications for fields such as affective computing, human-computer interaction, and artificial intelligence. This thesis explores the challenging problem of recognizing real/fake emotions in video data. Initially, a single-task learning approach is utilized to classify emotions as either genuine or posed, followed by the application of transfer learning and fine-tuning techniques to enhance their performance. Subsequently, the approach is extended to a multi-task learning framework by separating the main classification task into two: one for recognizing emotions and the other for determining the authenticity of each emotion. Furthermore, the study highlights the significance of temporal information in recognizing genuine versus posed emotions, as demonstrated by the literature review and the development of a video classification pipeline with temporal information. The results from each approach are thoroughly analyzed, and possible future research directions are presented. Overall, this study contributes to the development of effective techniques for real/fake emotion recognition, which can have practical applications in areas such as law, entertainment, and healthcare.

I would like to dedicate this master thesis to my parents for their unwavering support, guidance, and encouragement throughout my academic journey. They have always been my pillars of strength and motivation, and I am forever grateful for their love and sacrifices. I also want to dedicate this work to my girlfriend, who has been a constant source of inspiration, motivation, and support throughout my academic journey. Her love, patience, and encouragement have helped me through the difficult times, and her unwavering belief in me has pushed me to achieve greater heights. I cannot express enough gratitude for all that she has done for me. To my parents and my girlfriend, I dedicate this work as a testament to the love and support that has made my achievements possible. Thank you for being my rock and my inspiration, and for always believing in me.

Chapter 1

Introduction

Facial expressions are an essential part of human communication. According to Dr Albert Mehrabian's extensive research[6] on the topic of body language, resulting in the 7-38-55 per cent rule. This rule indicated that only 7% of all communication is done through verbal communication, whereas the non-verbal component of our daily communication makes up 38% of the speaker's body language and 55% of facial expressions. Facial expressions are controlled by a complex network of muscles that are activated by the limbic system in response to emotional stimuli[21]. For example, when people experience a positive emotion like happiness, the facial muscles will typically move to create a smile, whereas when experience a negative emotion like anger, the facial muscles will move to create a scowl. Overall, facial expressions are a powerful means of nonverbal communication that can convey a wealth of information about a person's emotions, intentions, and social cues, making them an important aspect of human communication and interaction.

Emotion recognition and facial expressions are closely correlated since the latter is one of the most important and reliable indicators of a person's emotional state[26]. Furthermore, emotion recognition is an important source of emotional and social information in interpersonal communication. Knowing what other individual feels is relevant in predicting a person's psychological state, such as future behavior and the outcome of social interactions[27]. Recognizing emotions from facial expressions generally could be done easily by individuals. However, determining the authenticity of such expressions is a difficult task. Subtle differences between genuine and posed facial expressions are very small facial details that are hard to recognize by the human eye. Currently, limited deep learning research has dealt with identifying fake emotions with accuracy in a range of 51-76%. The minor differences between genuine and posed emotions make their detection a complex problem. Hence, various deep learning model architectures should be investigated using dedicated loss function and training strategies. During my master thesis on the detection of genuine and posed emotions, a video dataset will be explored and used to make experiments with several model combinations to develop a deep learning model for the proposed task.

In the world of deep learning, researchers and developers have used facial recognition technology to create algorithms that can detect and analyze facial expressions in real-time to determine a person's emotional state[61]. These algorithms use machine learning techniques to identify specific patterns of facial muscle movements associated with different emotions. This technology has numerous applications, including in healthcare, marketing, and law enforcement. For example, it can be used to diagnose and treat mental health disorders, analyze consumer responses to advertisements, and identify individuals who may be engaging in criminal activity. Moreover, emotion recognition has known a broad interest over the past two decades and especially facial

expression recognition, which represents the central axis the researchers explored. As same as image processing, facial expression recognition achieved a point of maturity due to two main factors: The availability of massive databases and the significant increase in computing power with GPU which both allowed the use of sophisticated algorithms, such as deep learning[32].

Emotions are an essential aspect of human communication and interaction, and the ability to accurately recognize them is crucial for various domains such as psychology, education, entertainment, and law. However, detecting genuine emotions from fake ones is a challenging task that requires both high-level cognitive and perceptual abilities. The main causation for that is humans are very skilled in concealing their true affective states from others. Even for professional psychologists, it is difficult to recognise deceit in emotional displays as numerous factors need to be considered[42, 40]. Many potential applications could benefit from the ability to automatically discriminate between fake and real facial emotions. For instance, improved human-computer interaction, improved human-robot interaction for assistive robotics[8, 33], treatment of chronic disorders[37] and assisting investigation conducted by police forces[1, 41].

More specifically, the ability to accurately detect real and fake emotions has significant social implications across various domains. In psychology, the detection of fake emotions can help diagnose psychological disorders such as sociopathy and borderline personality disorder, which are often characterized by the inability to express or recognize genuine emotions. Furthermore, the ability to identify real emotions can help therapists and counsellors to establish rapport and trust with their clients, leading to more effective therapy sessions. Additionally in education, the detection of real and fake emotions can aid in the development of more personalized learning environments. For instance, an intelligent tutoring system can use emotion recognition to detect when a student is frustrated or disengaged and provide additional resources or personalized feedback to keep the student motivated and engaged in the learning process. In the entertainment industry, the detection of real and fake emotions can enhance the immersive experience of video games and virtual reality environments. For instance, an avatar in a game can respond differently depending on whether the player's emotions are genuine or fake, leading to a more realistic and engaging experience. Moreover, in the field of law enforcement, the detection of fake emotions can be useful in detecting deception, which is crucial in criminal investigations. The ability to accurately detect lies can help prevent false convictions and reduce wrongful imprisonment. Overall, the detection of real and fake emotions has significant social implications and has the potential to improve various aspects of human communication and interaction. By developing a robust and accurate system for emotion recognition, we can pave the way for more advanced applications in affective computing, which can revolutionize how humans interact with technology.

With the advancement of deep learning techniques, researchers have explored various methods to address this problem. For example, LSTM[24] networks and SVMs[49] are used in several approaches combined with other models to tackle the fake and real emotion recognition task. Most of those approaches tackle the problem as a single task problem, i.e. they are trying to classify the emotion and the authenticity of it at the same time as a single task. In contrast with previously mentioned approaches, this work aims to leverage the Multi-Task Learning ability for the detection of real and fake emotions and compare it with single-task and spatial-temporal information approaches.

Multi-label, Multi-Task Deep learning is a type of deep learning approach that can handle multiple tasks and multiple labels in a single model[46]. In Multi-label classification, a single instance can be assigned with multiple labels, while in traditional binary or multi-class classification, a single instance can only be assigned with a single label. For example, in an image classification task, a single image can contain multiple objects, so each object can be assigned a label, resulting in a multi-label classification problem. In Multi-Task learning, the model is

trained to perform multiple related tasks simultaneously, where the tasks are related and sharing some common features. This allows the model to learn a shared representation that can improve the performance of all tasks. For example, in computer vision[54], a single model can be trained to perform multiple tasks such as image classification and segmentation at the same time. The Multi-label, Multi-Task Deep learning approach combines both Multi-label and Multi-Task learning into a single model. It is used to handle multiple tasks and multiple labels in a single deep learning model. This approach can be useful in various applications where multiple tasks and multiple labels need to be handled simultaneously.

In this work, the concept of multi-task learning is utilized to address the problem as two classification problems. In practice, the model is trained to classify the emotion of a given image as a first task and if it is real or fake as a second task simultaneously, which can be applied to various real-world scenarios such as detecting deception, emotion recognition in social media, and human-robot interaction. This research contributes to the field of artificial intelligence by exploring the potential of multi-task and multi-label deep learning approaches in emotion recognition, which has significant implications for human-computer interaction and affective computing. In this study, the SASE-FE[31] which is the first dataset of genuine and deceptive facial expressions of emotions will be used.

Furthermore, an objective of the present master thesis is to answer the following research questions

1. How could a state-of-the-art architecture on emotion recognition be modified to capture the subtle differences between real and fake emotions?

2. What is the effect of the multi-label focal loss function compared with the standard loss functions on the recognition task?

3. How important is the temporal aspect in capturing the differences between real and fake emotions and which is the best-suited temporal architecture for this task?

Chapter 2

Literature Review

Emotion recognition in videos is a challenging task due to several factors, including variations in facial expressions, lighting conditions, head poses, occlusions, and non-uniform backgrounds. Unlike images, videos contain temporal information that requires capturing and processing the dynamics of facial expressions over time. Furthermore, individuals can display different emotional states in a single video, making it challenging to identify the dominant emotion.

A solution[4] was presented to the Emotion Recognition in the Wild 2016 Challenge[13] in the video-based emotion recognition category, which aims to identify human emotions through facial expressions, voice, body language, and other cues. The proposed system focuses on recognizing emotions in video streams from trimmed clips and can predict one of seven emotion labels, including the six basic emotions (Anger, Disgust, Fear, Happiness, Sad, Surprise) and Neutral. The system consists of several modules, including face detection, image preprocessing, deep feature extraction, feature encoding, and an SVM classifier[10]. The proposed system[4] achieved a validation accuracy of 59.42%, surpassing the competition baseline of 38.81%. On the test data, the system achieved a recognition rate of 56.66%, also improving the competition baseline of 40.47%.

In more detail, the proposed approach[4] for emotion recognition from videos using images is a pipeline of several modules that work together to achieve high accuracy in emotion recognition. The face detection module uses a Haar cascade classifier to detect faces in each video frame. The detected face images are then preprocessed by cropping and resizing to a fixed size, followed by histogram equalization and normalization to enhance their quality. The deep feature extraction module uses a pre-trained deep convolutional neural network (CNN) to extract high-level features from the preprocessed face images. Specifically, the last fully connected layer of the VGG-Face network is used as a feature extractor. The feature encoding module uses a spatial pyramid pooling (SPP)[19] method to encode the extracted features into a fixed-length vector. This allows the system to handle variable-length input sequences. The SPP layer partitions the input feature maps into multiple spatial bins and pools the features within each bin, resulting in a fixed-length feature vector. Finally, the SVM classifier is trained on the encoded feature vectors to predict the emotion label of the video sequence.

In conclusion, the proposed approach for emotion recognition in the wild from videos using images is a well-designed pipeline of several modules that achieve high accuracy in recognizing emotions. The system significantly outperforms the competition baseline, showing the effectiveness of the proposed approach.

An interesting approach to tackle the problem of discriminating between genuine and fake

emotions proposed a new model by combining a mirror neuron modelling and deep recurrent networks[24], called long-short term memory (LSTM)[22] with parametric bias (PB), by which features are extracted in the spatial-temporal domain from the facial landmarks, and result in two PB vectors, one for genuine and other for fake one. Additionally, a binary classifier based on gradient boosting[3] was used to enhance the discrimination capability between two PB vectors. That approach achieved a 66.7% accuracy in the SASE-FE dataset[31] benchmark.

The core architecture of the proposed system was designed to integrate the advantages of RNN-PB and LSTM. It keeps the long-term dependency and yet exploits the powerful discrimination capability of LSTM against the subtle difference between fake and genuine facial expressions. The system detects the face in the first frame of a test video and then tracks this face in the remaining frames. Then, extracts facial landmarks and then learns parametric bias (PB) vectors from a testing stream. These PB vectors are classified using the gradient boosting machine. The LSTM-PB and the gradient boosting machine are trained using the challenge dataset and their results are placed as the 1st place of the ChaLearn 2017 challenge[59]. More specifically in this approach[24], firstly, they combined a Haar-feature face detector[58] and a MOSSE-based object tracker within the OpenCV[7] environment to detect the faces in the dataset videos. Secondly, the facial landmarks are detected using the DLib library[29], which implements an ensemble of regression trees for detecting landmarks. Then, they introduced the long-short memory with parametric bias LSTM-PB, which learns time sequences in a supervised manner. The only difference is that the 2D-Grid LSTM is used to store memory information during the learning process. The backpropagation through time (BPTT) algorithm[60] is utilized in training the structural properties of the training time sequences. Meanwhile, PB vectors encode the specific properties of each time sequence simultaneously. As a consequence of the learning process, the LSTM-PB self-organizes a mapping between PB vectors and time sequences. Finally, to learn the PB vectors, they utilized a variant of the BPTT algorithm.

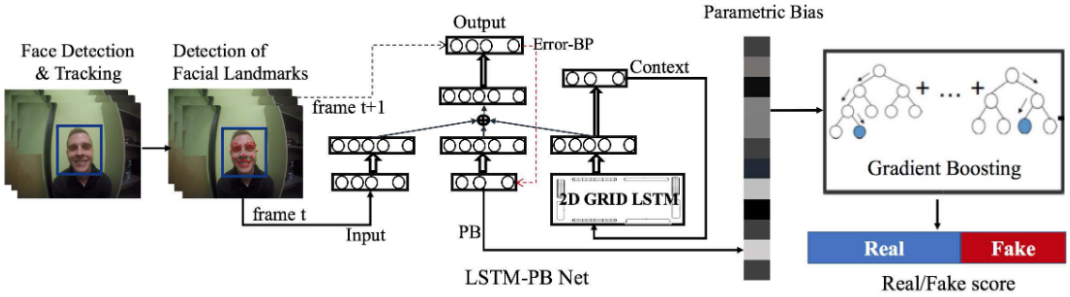


Figure 2.1: The pipeline of the framework for real versus fake expression recognition. Figure adapted from the original paper "Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks"[24].

Overall, the proposed model is representative of the dynamical system and has three operational modes: learning, generation and recognition modes. During the training of the LSTM-PB, two parametric biases are created, one for the genuine emotion and one for the fake one. The goal of the training is to update weight sets such that the network becomes a time series predictor for the facial stimuli and to create the two PB vectors. Then, in the generation mode, the network produces a stream of facial landmarks corresponding to either a genuine or a fake facial expression, depending on the given PB vectors. In the recognition mode, LSTM-PB observes the given stream of facial landmarks and computes a PB vector that matches with the pre-trained one.

When the network makes an initial prediction, the error between the prediction and the target is generated at the output layer. Then, the prediction error is back-propagated to the PB vector in terms of mean square error. If pre-learned facial landmark movement patterns are perceived, the PB values tend to converge to the values that have been determined in the learning phase. Lastly, the final classification is done with Gradient Boosting. Overall, the contributions of the paper include, the proposed LSTM-PB model as a result of the combination of mirror neuron modelling with an extra addition of a gradient boosting classifier, and lastly the first approach of that type of model to solve a classification problem recognizing fake facial expressions.

Another interesting approach, to tackle the real vs fake emotion recognition problem proposed the Enhanced Boosted Support Vector Machine algorithm (EBSVM)[49] for determining important thresholds required to understand fake emotions. Also, they created a new dataset named FED with real and fake emotion images and used them with experiments along with SASE-FE[31]. The HaarCascade algorithm is utilized to detect faces in input images. The method of parallel cascade linear regression (ipar-CLR)[2] is applied to locate landmarks. The 92 fiducial points are identified from various facial components. Then, to address the real and fake emotion recognition challenge, the authors propose a novel technique called the Enhanced Boosted Support Vector Machine (EBSVM) algorithm. The EBSVM algorithm determines important thresholds required to understand fake emotions and considers the entire data for classification at each iteration using the ensemble classifier. The EBSVM algorithm consists of two stages: feature selection and classification. In the feature selection stage, the authors use a novel method called the ReliefF algorithm[53] to identify the most relevant features for discriminating between real and fake emotions. ReliefF works by estimating the relevance of each feature based on its ability to discriminate between the two classes while taking into account feature interactions. In the classification stage, the authors use an ensemble of SVM classifiers to classify the images as either real or fake emotions. The ensemble classifier combines the outputs of multiple SVM classifiers, each trained on a random subset of the feature space. The authors use the AdaBoost algorithm[50] to weight the SVM classifiers according to their classification accuracy, with the more accurate classifiers being assigned higher weights. The EBSVM algorithm achieved 98.08% classification accuracy for different K-fold validations, indicating a significant improvement in detecting fake emotions compared to previous methods. The authors conclude that their proposed EBSVM algorithm could be used in applications that require accurate differentiation between real and fake emotions. The contributions of the paper are the proposal of the Enhanced Boosted Support Vector Machine (EBSVM) Algorithm, the construction of the FED dataset and also contribute to the identification of key features that constitute real and fake emotions. Overall, the authors' proposed approach of using a novel feature selection algorithm and an ensemble of SVM classifiers shows promise in accurately differentiating between real and fake emotions in facial expression recognition and emotion detection.

The paper "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks"[48] explores the use of lightweight convolutional neural networks (CNNs) for facial expression and attribute recognition, including age, gender, and ethnicity. This approach focuses on multi-task learning, where a single model is trained to perform multiple tasks simultaneously, as a way to improve performance and reduce the computational cost of training and inference. More specifically, they train and evaluate several models based on MobileNet[23], EfficientNet[56], and RexNet[18] architectures, which are designed to be lightweight and efficient for mobile and embedded applications. The models are trained on cropped faces without margins, which reduces the size of the input and focuses on the facial features that are most important for recognition. The authors also highlight the importance of fine-tuning the models

to predict facial expressions, which requires more complex and nuanced features. Furthermore, the models are evaluated on several datasets, including the UTKFace[62] dataset for age, gender, and race recognition, and the AffectNet dataset for emotion classification. The authors show that their models achieve near state-of-the-art performance on these datasets, demonstrating the effectiveness of their approach. For example, their models achieve an accuracy of 98.5% for gender classification and 94.2% for age classification on the UTKFace dataset, and an accuracy of 67.3% for emotion classification on the AffectNet dataset[38]. The authors also demonstrate the usefulness of their models as feature extractors for facial regions in video frames. They show that using the trained models as feature extractors leads to improved accuracy compared to previously known state-of-the-art single models for the AFEW and VGAF datasets[30, 13] from the EmotiW challenges, which focus on facial expression recognition in video sequences. Finally, the paper presents a comprehensive study of multi-task learning of lightweight CNNs for facial expression and attribute recognition. Lastly, this work demonstrates the effectiveness of their approach on several datasets and shows that their models can be used as feature extractors for video-based recognition tasks.

An influence approach for this work is the paper "Facial Emotion Recognition: A multi-task approach using deep learning"[47]. This paper proposed a multi-task learning algorithm, in which a CNN detects the gender, age and race of the subject along with their emotion. The validation of this method was done using two datasets containing real-world images, resulting in significantly better accuracy than the state-of-the-art algorithms for this task. The proposed multi-task learning approach includes a single model generating labels for emotion, gender, age and race compared with single-task learning approaches. For that purpose, each of the classification tasks has a separate output layer, which means that each output layer has a softmax function which is independent of the softmax function of other output layers. For emotion recognition, the model classifies the input image into one of 7 classes consisting of basic human emotions. The gender is classified as male, female or unsure and race as Caucasian, African-American or Asian. For age estimation, the classes are divided into 5 ranges of ages. The datasets used are FER containing labels for only emotion and RAF-Db[35, 34] containing labels for all corresponding tasks. In the pre-processing stage, the images have been subjected to pose normalisation. This has been achieved by identifying the eye centres of the faces presented in the images, and the images being rotated such that the line joining the eye centres becomes horizontal. The eye detection has been carried out using the Cascade classifier of the OpenCV library[7]. There were a few cases in which the pre-trained algorithm was incorrectly identifying the eye centres and the images were being rotated by an angle of large magnitude. The CNN structure was heavily influenced by previous work, but some modifications were made to enhance accuracy by adjusting the dropout rates. The previous study used dropout rates of 0.4, 0.4, 0.5, and 0.6 in the increasing layers of the neural network. However, in the current study, the dropout rates are decreasing in a monotonous manner with 0.6, 0.5, 0.4, and 0.4 in the increasing layers of the neural network. As a result, the accuracy of the model has improved from 63% to 67% when tested on the FER dataset[63] using cross-validation. For each of the classification tasks, categorical cross entropy is used to calculate the loss of each task. Consequently, there are four loss functions, each corresponding to a specific task. To obtain a significant outcome for each task and for the entire Neural network, it is essential to allocate appropriate weights to each loss function. The weights assigned to each loss function for emotion, age, race and gender were 2, 4, 1.5, and 0.1, respectively. After determining the weights, the overall loss is computed, and the model backpropagates this loss to refine its performance in subsequent iterations. To guarantee that the CNN achieves optimal performance, two callbacks were implemented: Early stopping and Reduce on Plateau. When using multi-task learning, the validation accuracy is

79% when training the model with all the labels on RAF-Db. However, when the same model is pre-trained on RAF-Db and then trained solely on FER, which only has labels for emotions, the accuracy drops significantly to 53%. As a result, the best-performing model for multi-task learning was trained solely on RAF-Db. In summary, this study contributes a new approach to multi-task learning, which involves predicting gender, age, race, and emotion simultaneously. The results obtained using this approach were significantly better than the traditional single-task learning approach using the same CNN architecture. This approach can be applied to various CNN architectures for facial emotion recognition. To train more data effectively and obtain all labels, pre-trained open-source models for the FER dataset can be used to generate labels for age, gender, and race and incorporate them into the training set. Additionally, various filters such as Gabor, HOG, LBP, and SIFT can be utilized during the preprocessing step to assess their impact on the results. Finally, rather than hardcoding values, the optimal weights of the loss function for the four branches of the CNN can be determined.

Chapter 3

Methodology

3.1 Single-Task Learning Approach

Single-Task Learning (STL) is a machine learning technique in which a model is trained to perform a single specific task. In the context of computer vision and deep learning, a single task may include image classification, object detection, semantic segmentation, or any other computer vision task that involves analyzing visual data. The objective of STL is to optimize a single loss function that measures the difference between the predicted output of the model and the ground truth output. The loss function is typically chosen based on the task at hand, and it can take different forms, such as cross-entropy loss, mean-squared error, or binary cross-entropy, among others. STL has been widely used in computer vision applications and has achieved impressive results in a variety of tasks. However, one of the limitations of STL is that it does not take into account the fact that many computer vision tasks share common features and patterns. This means that even if two tasks are related, they are often trained separately, which can lead to inefficiency and non-optimal results. For example, consider the task of image classification and object detection. In image classification, the goal is to classify an entire image into one of several categories, whereas, in object detection, the goal is to detect the presence and location of one or more objects in an image. These two tasks share many common features, such as edges, corners, and texture, which could be learned jointly in a multi-task learning framework. Despite its limitations, STL remains a valuable and widely used approach to solve many computer vision problems. It is simple to implement and can achieve high accuracy if the task is well-defined and the training data is representative of the target domain. Moreover, it provides a baseline to compare the performance of more complex methods, such as multi-task learning, transfer learning, or domain adaptation.

In the present work for the single-task learning approach, the SASE-FE dataset[31] prepared and organized to 12 main labels of real and fake emotions. Those are: fake_surprise, fake_angry, fake_contempt, fake_disgust, fake_sad, fake_happy, real_angry, real_contempt, real_disgust, real_happy, real_sad, real_surprise. The frames that are generated from each video duration range from 40% to 90% for each label, as these frames contain crucial information that each participant displays for each emotion. I have chosen not to take into account certain parts of the videos because they either display a natural or neutral facial expression, or occasionally a random emotion. In order to obtain significant information for the identification of real and fake emotions, the Haar Cascade classifier[44] was employed to detect the faces which were subsequently cropped to generate input processed images for the training process. The resultant images were then input into a baseline model for training purposes. The pipeline of the Single-task approach is

demonstrated in figure 3.1, and the baseline models in table 3.1.

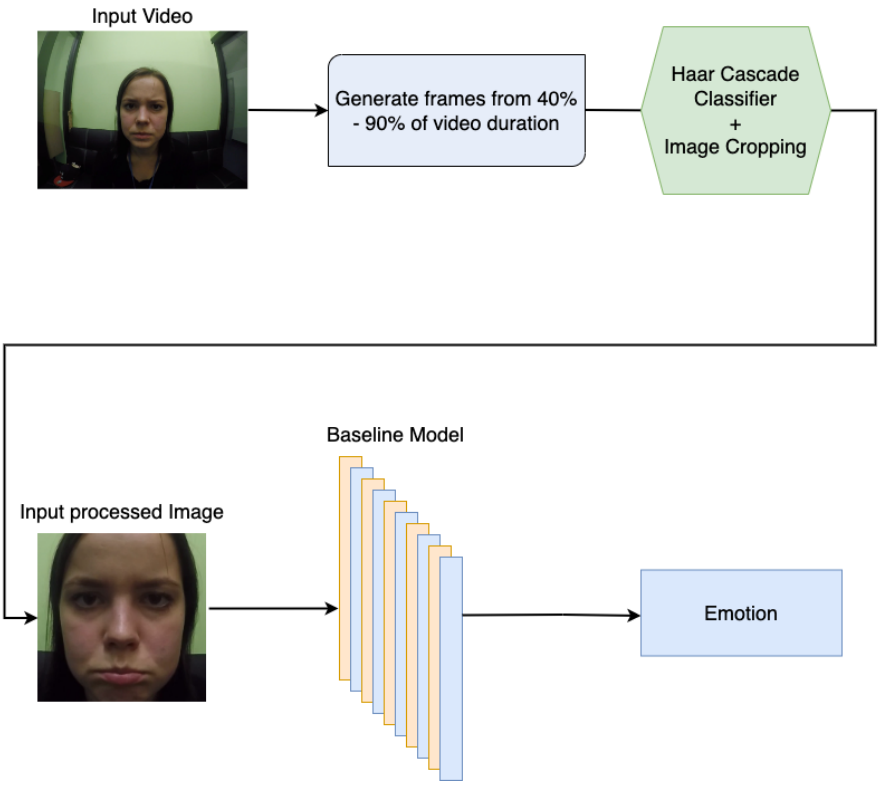


Figure 3.1: Single-Task learning approach pipeline.

The overall methodology of this approach includes the following steps:

1. Preparation of the data structure, define the labels.
2. Generation of frames from every video.
3. Data pre-processing: Face detection and cropping.
4. Training of a model with classification head.

Model name	**pre-trained weights**
VGG16[52]	ImageNet[12]
InceptionResNetV1[51]	VGGFace2[9]
EfficientNetV2M[57]	ImageNet[12]
ViT-b16[14]	ImageNet[12]

Table 3.1: Models and their pre-trained datasets used in single task frames approach.

The initial baseline models were, VGG16[52], EfficientNetV2M[57], and ViT-b16[14] which are all popular deep neural network architectures that have been pre-trained on the ImageNet[12] dataset, a large-scale image classification dataset with over a million labelled images.

VGG16 is a convolutional neural network architecture that was developed by the Visual Geometry Group at the University of Oxford. It has 16 layers, consisting of 13 convolutional layers and 3 fully connected layers. VGG16 is known for its simplicity and effectiveness in image classification tasks, achieving high accuracy rates on the ImageNet dataset. The architecture uses small 3x3 convolutional filters throughout the network, which helps to preserve spatial information while reducing the number of parameters. The final layer of the network is a softmax layer that produces a probability distribution over the classes. Despite its success, VGG16 has a large number of parameters, making it computationally expensive to train.

EfficientNetV2M is a more recent CNN architecture that was introduced in 2021. It is based on a novel compound scaling method that efficiently balances the number of model parameters, model depth, and image resolution, resulting in a highly efficient and accurate model. The architecture consists of a stack of convolutional layers, followed by a global average pooling layer, and finally a softmax layer. The convolutional layers use a combination of depthwise separable convolutions and standard convolutions, which helps to reduce the computational cost while maintaining accuracy. EfficientNetV2M has shown superior performance on various computer vision tasks, including image classification, object detection, and semantic segmentation.

ViT-b16, or Vision Transformer, is a transformer-based neural network architecture that was introduced in 2020. Unlike traditional convolutional neural networks, which rely on hand-designed feature extractors, ViT-b16 uses self-attention mechanisms to learn the feature representations directly from the raw input images. The architecture consists of a stack of transformer blocks, followed by a global average pooling layer, and finally a linear projection and softmax layer. Each transformer block consists of a multi-head self-attention mechanism and a feedforward network. ViT-b16 has shown impressive results on various image classification benchmarks, including the ImageNet dataset. However, it requires more computational resources than traditional CNNs.

All three of these architectures have achieved state-of-the-art performance on various image classification tasks, with VGG16 being a classic and highly interpretable model, EfficientNetV2M being highly efficient and accurate, and ViT-b16 being a novel transformer-based approach that achieves excellent performance. They have all been pre-trained on the ImageNet dataset, making them useful for transfer learning on a variety of computer vision tasks.

Hence, the next model that was used as a baseline was the InceptionResNetV1 introduced in the paper "FaceNet: A Unified Embedding for Face Recognition and Clustering"[51] proposes a novel approach to face recognition and clustering using deep neural networks. InceptionResNetV1 is a hybrid model that combines the Inception[55] and ResNet[20] architectures to achieve state-of-the-art performance in face recognition tasks. The architecture includes a stem module that preprocesses the input image, followed by multiple Inception-ResNet modules that contain a combination of Inception and ResNet blocks. The Inception-ResNet modules enable the model to capture both local and global features of the face. VGGFace2 is a large-scale dataset that contains over three million images of faces from various demographics and ethnicities. The dataset is used to train the InceptionResNetV1 deep neural network architecture for face recognition. The weights of the model are useful in facial expression recognition tasks, as the model has learned to capture high-level features of the face, which are relevant to facial expression recognition. These weights can be used to fine-tune an existing CNN model for facial expression recognition tasks or to initialize the weights of a new CNN model.

The architecture of InceptionResNetV1, with its inception modules and residual connections, is designed to be highly effective at extracting features from images. This allows the model to capture and utilize the subtle and complex facial expressions that are important for emotion recognition. VGG16, EfficientNetV2M, and ViT-b16 were designed for general computer vision tasks and may not be as effective at capturing the nuances of facial expressions. Also, the

pre-trained weights of InceptionResNetV1 on VGGFace2 provide a strong starting point for training the model on emotion recognition tasks. By starting with weights that have already been trained on a large dataset of faces, the model requires less fine-tuning to perform well on facial expression recognition tasks. Overall, InceptionResNetV1 pre-trained on VGGFace2 is a better choice for facial expression recognition because of its specialized training on facial recognition tasks, effective architecture for feature extraction, and pre-trained weights on a large-scale face recognition dataset.

3.2 Multi-Task Learning Approach

In computer vision and deep learning, Multi-Task learning (MTL) and Single-Task learning (STL) are two commonly used approaches to train neural networks. Single-Task learning is the traditional approach in which a model is trained to perform a single specific task, such as object recognition, image segmentation, or image classification. The network is optimized to minimize a single objective function, which is typically a loss function that measures the difference between the predicted and true values of the target variable. On the other hand, Multi-Task learning involves training a single model to perform multiple related tasks simultaneously. In MTL, the model is optimized to minimize multiple loss functions, each corresponding to a different task. For example, a single model can be trained to recognize objects in images, detect their poses, and segment them. By training a model to perform multiple tasks at once, MTL can improve the overall performance of the model on all the tasks compared to STL. This is because the model learns to share features across tasks and can use the information learned from one task to improve its performance on another. Additionally, MTL can reduce the computational cost of training multiple models separately for each task. However, designing an effective multi-task loss function can be challenging, and the model may overfit on some tasks or underperform on others. By jointly optimizing multiple loss functions, the model is encouraged to learn more robust features that generalize well to new data. However, designing an effective multi-task loss function can be challenging, as the different tasks may have different scales and importance, and balancing between them can be difficult. Moreover, the model may overfit on some tasks and underperform on others, especially if the tasks are not closely related. MTL has been successfully applied in many computer vision applications, including object detection, semantic segmentation, and pose estimation, among others. It has also been extended to other domains, such as natural language processing and speech recognition, where it has shown promising results.

Instead of facing the problem as a single task, i.e. Multi-class classification, in this approach, the problem of distinguishing real and fake emotions is treated as two tasks that are relying on common features. The first task is the multi-class classification for the emotion of a given input and the second is a binary classification of its genuineness (real or fake).

The first steps of this approach are similar to the single-task approach. Those include the frame generation from the 40%-90% of each video duration and then the face detection and cropping, resulting in frame images with participants' faces for each emotion.

The second step of the MTL pipeline includes the creation of a multi-task dataset structure in a format that the labels and the images are used respectively for each task. For convenience of this the image frames were renamed in the following format: [real/fake]_[emotion]_[frame_number].jpg. For instance, the image frame file named "1_4_100.jpg" is a frame from a person expressing the emotion of real happiness.

Subsequently, a baseline model, specifically InceptionResNetV1, was chosen and augmented with two additional heads positioned atop it. This decision was made after obtaining the results of the single-task approach and observing that a model pre-trained on an emotion-related dataset

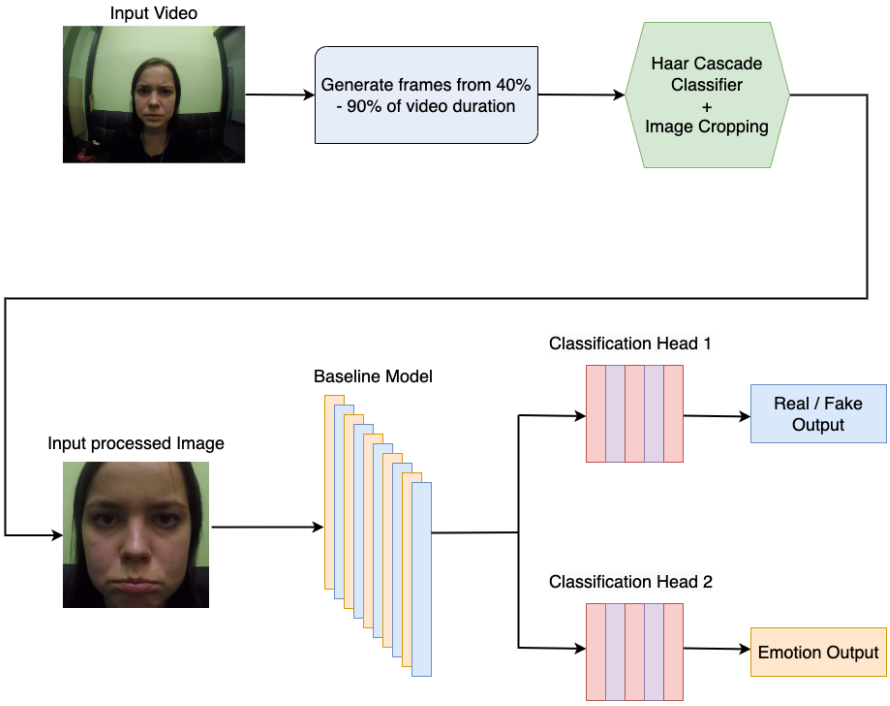


Figure 3.2: Multi-Task Learning Approach pipeline.

will be suitable to host the baseline of the multi-task model. However, experiments with a similar approach named HydraNet[39], were conducted for the evaluation of the final model.

Hence, two classification heads were incorporated into the InceptionResNetV1 model architecture to enable the model to perform two distinct tasks. The first head was comprised of a linear layer with ReLU activation, followed by another linear layer with two output units. The second head, on the other hand, consisted of a similar linear layer with ReLU activation, followed by another linear layer with six output units. These heads were implemented to enable the model to perform two tasks, namely binary classification and multi-class classification, respectively. An illustration of the MTL pipeline is demonstrated in figure 3.2.

3.2.1 Combined Accuracy

An important difference with the STL approach is the definition of the model's accuracy and loss. In the MTL approach, the accuracy of each task is calculated separately to get a better overview of how the model performs in each task. However, the overall model accuracy is calculated, to establish this approach comparable with other approaches and give an overview of the overall performance.

More precisely, first, the accuracy of the emotion recognition task is calculated by comparing the predicted emotion label with the ground truth emotion label and summing up the number of correct predictions. This is done for all samples in the test set. Then, the accuracy for the real/fake classification task in a similar way by comparing the predicted label with the ground truth label. Finally, the code calculates the overall accuracy by counting the number of samples for which both tasks are correctly classified (i.e., both predicted emotion labels and predicted

real/fake labels are correct). This is done by computing the element-wise logical AND between the two sets of predictions and summing up the number of True values. The result is divided by the total number of samples in the test set to obtain the overall validation accuracy.

3.2.2 Loss Function

In this work, two loss functions are mainly used for each task, namely the Cross-Entropy loss and Multi-Focal loss, and their combinations are reported in the results section 4.

Categorical Cross-Entropy

The categorical cross-entropy loss function calculates the loss of an example by computing the following sum:

$$\text{Loss} = -\sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i \tag{3.1}$$

where $\hat{y}_i$ is the $i-th$ scalar value in the model output, y_i is the corresponding target value, and output size is the number of scalar values in the model output.

Multi-Focal Loss

The multi-label focal loss (MFL)[5] is a modification of the binary focal loss[36] used for multi-label classification problems. It is designed to address the issue of class imbalance in the dataset and improve the model's performance. This loss function assigns different weights to each class based on their contribution to the overall loss. It down-weights easy examples and up-weights hard examples to focus the training process on the challenging examples, thereby improving the model's performance. The focal loss also introduces a tunable hyperparameter, the focusing parameter, which adjusts the degree of down-weighting or up-weighting of easy or hard examples. The multi-label focal loss is commonly used in object detection, image segmentation, and other multi-label classification tasks. The definition of the multi-label focal loss (MFL) as follows:

$$MFL_{\alpha,\gamma}(y,\sigma) = -\sum_{n=1}^{Ne} \alpha(1-\sigma_i)^{\gamma} y_i \log(\sigma_i) + (1-\alpha)\sigma_i^{\gamma}(1-y_i)\log(1-\sigma_i) \tag{3.2}$$

Where:

- Ne is the number of classes.
- σ_i and y_i denote respectively the categorical model output and its ground-truth label for the i^{th} class.
- α and γ are two empirical parameters.
  - γ, the focusing parameter, still has the same purpose as in the binary focal loss. It up-weights the hard classification samples while down-weights the easy ones, which in this case might be the less frequent ones.
  - On the other hand, α here does not serve to balance the two binary classes but to promote either recall or precision costs.

3.2.3 Combined Loss

The definition of the combined loss function is significantly more intricate than that of the accuracy metric. Previous methodologies have computed the total loss by summing the individual losses, which may not be optimal for a multi-task learning approach as one task may perform better and have a smaller loss, while the other may be inaccurate and have a larger loss. Hence, it is imperative to develop a more appropriate technique for defining a combined loss function in multi-task learning.

Furthermore, an alternative method for defining the combined loss is the weighted sum loss, where each task is assigned a weight, and the losses are then combined. However, it is challenging to determine the weight or importance of each task manually. For instance, determining whether correctly identifying emotions or distinguishing between genuine and posed, is more crucial and controversial. Therefore, the weighted sum loss is not the most suitable approach for calculating the combined loss in this scenario.

The primary notion was to employ the precision scores of each task as weights to compute the proposed combined precision loss. The precision scores for emotions and real/fakes represent the ratio of correct predictions made by the model out of all positive predictions for each category. Thus, in this methodology, the total loss was computed based on equation 3.3.

$$\text{Total\_Loss} = \text{precision\_emotions} \cdot \text{loss\_emotions} + \text{precision\_real\_fake} \cdot \text{loss\_real\_fake} \tag{3.3}$$

Overall, I think it's a better way to calculate the total loss of the model, because it is hard to give weight/importance to every task manually, the precision score can give such weight to each task to be used in the total loss calculation.

3.3 Temporal Learning Approach

Temporal learning is a crucial aspect of video classification, as it allows the model to analyze the content of a video over time and capture the temporal dependencies between the frames. In video classification, temporal learning can be achieved using various machine learning techniques, including Recurrent Neural Networks (RNNs) and 3D Convolutional Neural Networks (3D-CNNs).

3D-CNNs are a type of neural network that is designed to handle spatiotemporal data. Unlike traditional 2D-CNNs, which only consider the spatial information in an image, 3D-CNNs operate on a sequence of frames and can capture the temporal information between them. 3D-CNNs extend the concept of the 2D convolution operation to a third dimension, allowing them to extract spatial and temporal features simultaneously. The architecture of a 3D-CNN typically consists of multiple layers of 3D convolutional and pooling operations, followed by fully connected layers for classification. In the initial layers, the 3D-CNN extracts low-level features from the input frames, such as edges and corners. As the network goes deeper, the features become more abstract and complex, capturing higher-level patterns in the data. One advantage of 3D-CNNs over RNNs is that they can capture long-term dependencies in the video data efficiently. RNNs are known to suffer from vanishing gradients, which makes it challenging for them to propagate information across long sequences. In contrast, 3D-CNNs can model long-term dependencies more efficiently, as they can capture the temporal information in a larger window of frames.

There are several challenges in using 3D-CNNs for video classification. One challenge is the computational cost, as 3D-CNNs require more computational resources and memory than traditional 2D-CNNs. To address this issue, researchers have proposed various techniques, such

as group convolution and spatiotemporal factorization, to reduce the computational cost of 3D-CNNs. Another challenge is the limited amount of labeled video data, which makes it challenging to train deep 3D-CNNs from scratch. To overcome this challenge, researchers have proposed various transfer learning techniques, such as pre-training the 3D-CNNs on large-scale datasets, such as Kinetics 400[28], and fine-tuning them on smaller video classification datasets.

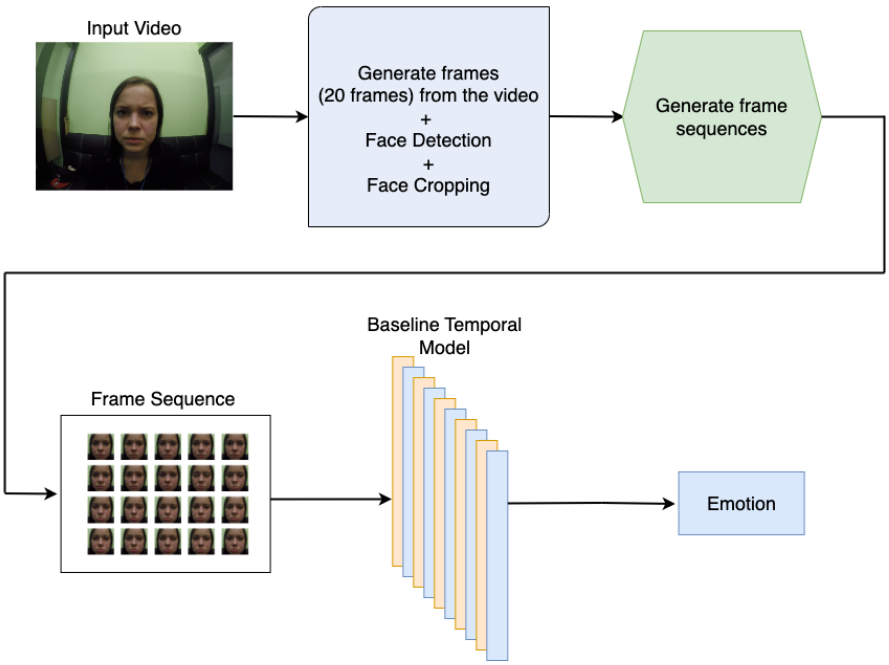


Figure 3.3: Temporal Learning Approach pipeline.

In the first step of the temporal approach, a given number of frames are created from each video, and they are pre-processed with Haar Cascade to detect participants' faces and crop them. Following, frame sequences that capture the temporal information for every video were generated. An example of such frame sequences is shown in figure 4.1. Then, a basic 3D-CNN model was trained within the produced frame sequences which classify an emotion among 12 classes (e.g. real_happy, fake_sad, etc.). Figure 3.3 demonstrates the temporal approach pipeline.

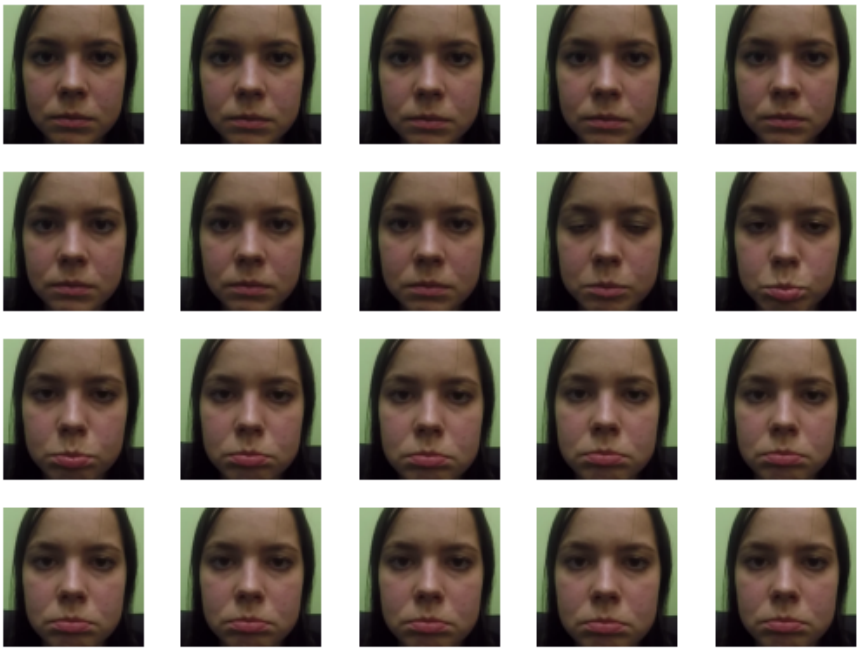


Figure 3.4: Example of a real_sad label frame sequence generated for the temporal approach.

Chapter 4

Results

4.1 Data

The SASE-FE dataset[31] created by the iCV Research Lab contains 600 different videos captured with high-resolution cameras recording 100 frames per second, containing video recordings of 50 participants of ages 19-36. For the experiments of this work, the data were split into 80% and 20% training and validation sets so that the same subjects were not present in both sets. The main reason behind the choice of such an age-range sample is that older adults have different, more positive responses than younger adults about feelings and emotions, and they are faster and more precise to regulate emotional states than younger adults[43, 11, 25]. More specifically, for each recording, participants were asked to act two facial expressions of emotions in a sequence, a genuine and a posed emotion. Genuine emotions are expressions of: happiness, sadness, anger, disgust, contempt, and surprise. In figure 4.1, samples from each emotion are shown and in table 4.1 summarized details for the SASE-FE dataset can be found.

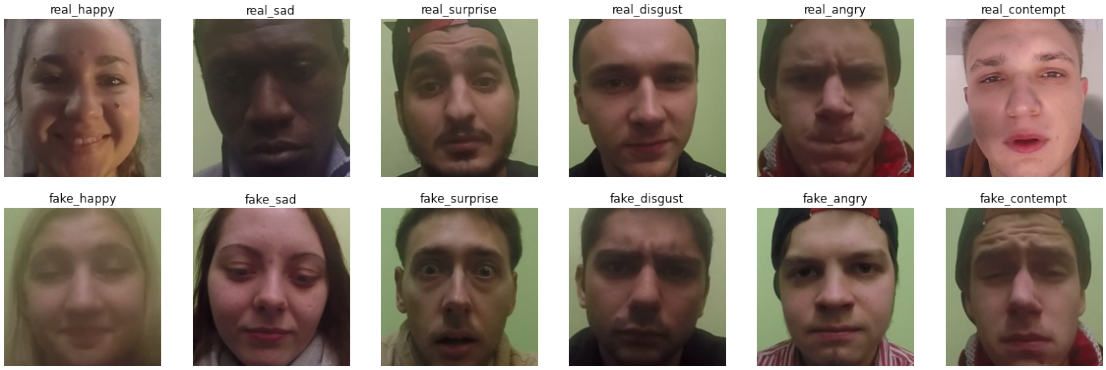


Figure 4.1: SASE-FE Database samples from real and fake emotions.

For eliciting genuine and realistic emotions, proposed videos based on emotion science research[17], were shown to the participants to increase the realism of their emotions. To increase the distinction between the two facial expressions presented in the sequence, the two emotions were chosen based on their visual and conceptual differences. Thus, the contrast was created by asking the participants to act happy after being sad, surprised after being sad, disgusted after being happy, sad after being happy, angry after being happy, and contemptuous after being happy. Note that the participants were asked to start their video recordings from a neutral face and none of the participants were aware of the fact that they would be asked to act with a second facial expression.

Dataset Name	SASE-FE
Source	iCV Research Lab (2017)
Expressions	6 emotions x (Real,Fake)
Subjects	50
# Videos	600
# Frames	100 fps
Male/Female	32/18
Ages	19–36
Format	MP4 - videos

Table 4.1: Summarized details of the SASE-FE dataset.

4.1.1 Data Augmentation

Data augmentation is a technique that can be used to artificially expand the size of a training set by creating modified data by using filters and computer vision techniques, from the existing dataset. It can be a beneficial method when there is not much diversity in the data. For instance, when we don't have many different angles of an object in a dataset. Also, data augmentation can be beneficial in achieving better generalization performance from a pre-trained model. In this approach, data augmentation techniques are being utilized including: random horizontal flip and random rotation of the images.

4.2 Evaluation Metrics

4.2.1 Confusion Matrix

A confusion matrix is a commonly used evaluation metric for assessing the performance of a classification model. It provides a summary of the model's correct and incorrect predictions in terms of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) rates. These rates can be used to calculate various metrics such as accuracy, precision, recall, and F1 score, which help to better understand the strengths and weaknesses of a model's performance.

The confusion matrix is represented as a table with four quadrants, where each quadrant corresponds to the number of instances that were predicted as positive or negative, and the actual outcome was positive or negative. For example, the top-left quadrant represents true positive predictions, i.e., instances that were correctly classified as positive, whereas the bottom-right quadrant represents true negative predictions, i.e., instances that were correctly classified as negative. The other two quadrants represent incorrect predictions, where false positives are

instances that were incorrectly classified as positive, and false negatives are instances that were incorrectly classified as negative.

In summary, the confusion matrix provides a visual representation of the model's performance, which helps to identify areas for improvement and to make informed decisions about adjusting the model's parameters or using different types of models.

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Here's an example of a confusion matrix:

		Predicted Label	
		Positive	Negative
True Label	Positive	TP	FN
	Negative	FP	TN

Table 4.2: Confusion Matrix example. Where: TN = True Negative, FP = False Positive, FN = False Negative, and TP = True Positive.

4.2.2 Accuracy

Accuracy is one metric for evaluating classification models. Accuracy is the proportion of the number of correctly classified data instances over the total number of data instances. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4.1}$$

Another equation to calculate the accuracy based on the confusion matrix shown in table4.2 is the following:

$$\text{Accuracy} = \frac{\text{TN}}{\text{TN+FP+TP+FN}} \tag{4.2}$$

4.2.3 Precision

Precision is a metric used to evaluate the performance of a classification model. It is the ratio of true positive predictions to the total number of positive predictions made by the model. In other words, precision measures the proportion of positive predictions that are actually correct. A high precision score indicates that the model is making few false positive predictions.

$$\text{Precision} = \frac{\text{Positive Predicted Value}}{\text{Total positive predicted data instance}} \tag{4.3}$$

Another equation to calculate Precision, based on the confusion matrix shown in table 4.2 is the following:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.4}$$

4.2.4 Learning Rate Scheduler

ReduceLROnPlateau: Reduce the learning rate when a metric such as accuracy or loss has stopped improving. Models often benefit from reducing the learning rate by a factor of 2-10

once learning stagnates. This scheduler reads a metrics quantity and if no improvement is seen for a number of epochs, the learning rate is reduced. Factor by which the learning rate will be reduced. $new\_lr = lr * factor$. Experiments made with factor equals 0.1 and 0.5.

4.3 Single-Task approach results

The outcomes of the single-task methodology encompass a comparison of the baseline models that were fine-tuned in the SASE-FE dataset. A classification head was introduced on top of each baseline model for the task. During some experiments, the networks were trained, whereas in other cases, the classifier was the only component that was trained while the baseline model was frozen. Our observations revealed that this approach was not beneficial as the ImageNet and VGGFace2 datasets were much larger than the SASE-FE dataset, resulting in overfitting to the models and inadequate outcomes. Consequently, training the entire network was selected, which yielded superior results. In all experiments, the cross-entropy loss function and stochastic gradient descent optimizer were used as standard choices to commence experimentation, while a reduction in loss plateau was applied to improve the models' performance by scheduling the learning rate.

Model	**Training Config**	**Loss Function**	**Optimizer**	**Weights**	**Loss**	**Accuracy**
VGG16	Train all	CE	SGD	-	9.029	15.922
VGG16	Train all	CE	SGD + LR-Scheduler	ImageNet	6.363	20.439
VGG16	Train all	CE	SGD	ImageNet	5.944	26.010
EfficientNetV2M	Train all	CE	SGD	ImageNet	5.373	26.644
VIT_B_16	Train all	CE	SGD	ImageNet	6.104	21.810
InceptionResNetV1	Freeze baseline + Train Classifier	CE	SGD	VGGFace2	2.315	19.892
InceptionResNetV1	Freeze baseline + Train Classifier	CE	SGD + LR-Scheduler	VGGFace2	2.335	18.497
InceptionResNetV1	Train all	CE	SGD	VGGFace2	4.569	27.817
InceptionResNetV1	Train all	CE	SGD + LR-Scheduler	VGGFace2	**4.655**	**29.062**

Table 4.3: Single-task approach experimental results.
Where CE: Cross-Entropy Loss, SGD: Stochastic Gradient Descent.

The VGG16, EfficientNetV2M, and ViT_B_16 models were pre-trained on ImageNet weights, while the InceptionResNetV1 model was pre-trained on VGGFace2, as also specified in section 3.1. The evaluation revealed that the ImageNet pre-trained models incurred higher loss without any significant accuracy score. The primary reason for this was that these models were trained to perform image classification in a large-scale generic dataset that lacked useful information regarding facial expressions and emotions. In contrast, the InceptionResNetV1 model, which was pre-trained on VGGFace2, appeared to be an optimal baseline for the emotion recognition task, achieving significantly better loss and accuracy scores. The experimental results from the single-task approach are summarized in Table 4.3.

The confusion matrix from the best results obtained in the single-task learning approach with the InceptionResNetV1 model is shown in Figure 4.2. One main observation from the confusion matrix is that the model is imprecise in recognising real and fake emotions. For example, the model is confusing the real_surprise with fake_surprise, resulting in a higher normalized score in the confusion matrix. Lastly, the model accuracy and loss during training are monitored and plotted in figures 4.3a and 4.3b respectively.

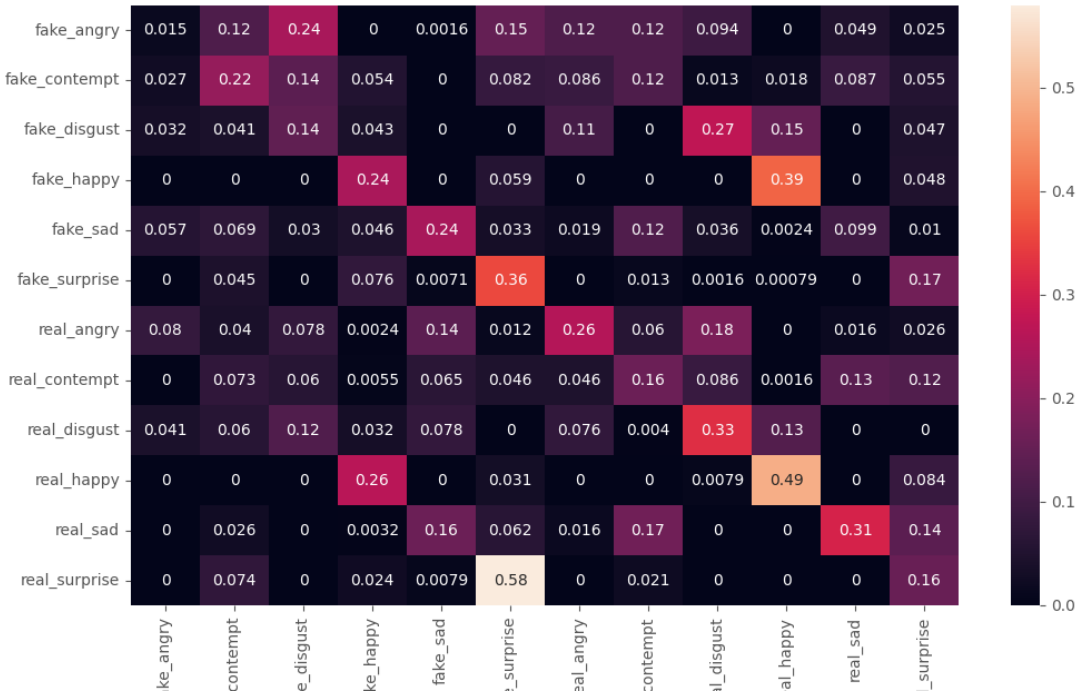
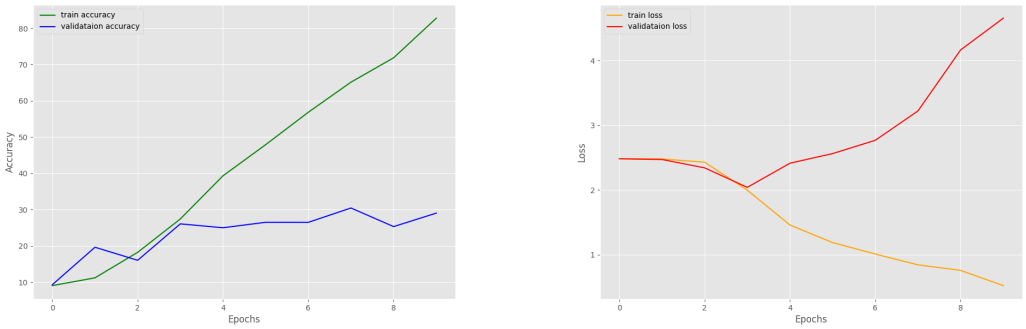


Figure 4.2: Confusion matrix of InceptionResNetV1 best results.



(a) Accuracy plot.

(b) Loss Plot

Figure 4.3: Best performing InceptionResNetV1 configuration accuracy and loss plots.

4.4 Multi-Task approach results

The first experiment conducted in the multi-task approach was about the selection of the baseline model. The table 4.4 presents a comparison of HydraNet and InceptionResNet models. These models share the same training configuration, including the use of cross entropy loss functions for

emotions and real/fake classifications, and optimization through stochastic gradient descent. The primary difference between the two models lies in their pre-trained weights. The data presented in the table shows that InceptionResNetV1 outperforms HydraNet in both emotion and real/fake classification accuracy. This suggests that InceptionResNetV1 provides a superior baseline for subsequent experiments.

Model	**Weights**	**Emotion Acc**	**Real/Fake Acc**	**Loss**
HydraNet	VGGFace2	42.593	48.387	2.147
InceptionResNetV1	ImageNet	53.721	55.322	2.787

Table 4.4: HydraNet vs InceptionResnetV1.

Moreover, another set of experiments includes the model performance comparison based on the training configuration and data transformations. The table 4.5 presents the results of training all part of the network or only the classifiers and using data augmentation and normalization or not. Two main observations arose from those experiments. First, freezing the baseline model and training only the classifiers did not perform well, since the data are not abundant. In contrast, when the complete network has been trained the emotion and real/fake accuracy increased. Second, it is noticed that when data augmentation is not used, the overall results are improved. The reasoning for that is that the data are not enough to include augmented images with filters. Eventually, when training the network without data augmentation better overall performance is achieved.

Model	**Train Config**	**Data Transformations**	**Loss**	**Emotion Acc**	**Real/Fake Acc**
InceptionResNetV1	Freeze Baseline + Train classifiers	DA + Normalization	2.329	33.217	48.284
InceptionResNetV1	Train all network	DA + Normalization	2.971	51.708	53.206
InceptionResNetV1	Train all network	Normalization	2.787	53.721	55.322

Table 4.5: Training configuration and data augmentation experiments. Where: DA is data augmentation as mentioned in section 4.1.1.

The experiments in question pertain to the selection of loss functions for emotions, real/fake classifications, and combinations of both. In particular, the use of the Multi-focal loss function (MFL), as described in Section 3.2.2, was investigated. These experiments were conducted using a uniform training configuration, which involved training the entire network without data augmentation, but with normalization. The SGD optimizer with a learning rate scheduler was used across all experiments. Further details of the results can be found in Table 4.6.

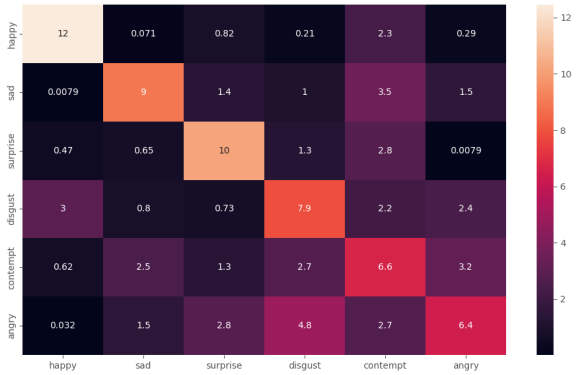
Model	**Emo. Loss**	**R/F Loss**	**Combined Loss**	**Emo. Acc**	**R/F Acc**	**Overall Acc**	**Overall Loss**
InceptionResnetV1	MFL	MFL	Precision Loss	54.609	53.713	30.150	1.265
InceptionResnetV1	MFL	MFL	Sum Loss	51.145	53.888	27.376	1.269
InceptionResnetV1	CE	CE	Precision Loss	49.227	55.893	28.034	1.588
InceptionResnetV1	CE	MFL	Precision Loss	51.145	54.411	26.536	1.781
InceptionResnetV1	MFL	CE	Precision Loss	50.892	52.318	28.192	1.870
InceptionResnetV1	CE	CE	Sum Loss	53.721	55.322	27.812	2.787

Table 4.6: Multi-task approach loss functions combinations results.

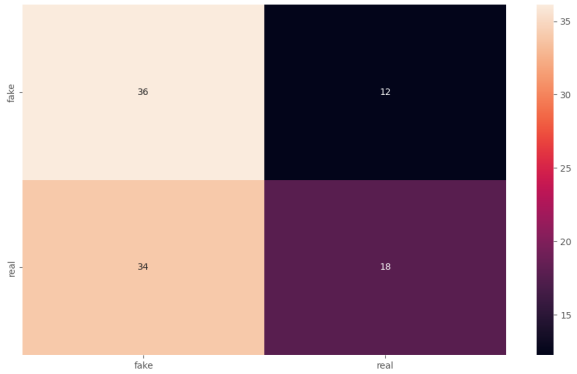
From Table 4.6, it can be observed that the MTL model performs better when trained with a combination of MFL loss functions for both emotion and R/F recognition tasks. When trained with this combination, along with precision combined loss, the model achieves an overall accuracy of 30.150% and an overall loss of 1.265. On the other hand, the model trained with the CE loss functions combination for both classification tasks, with combined sum loss, resulted in an overall

accuracy of 27.812% and an overall loss of 2.787. As shown in the table, the MTL approach can be useful in recognizing emotions and detecting whether they are real or fake. The results demonstrate the importance of selecting the appropriate loss function combination to train the model for achieving better performance in the given tasks.

Moreover, the confusion matrices for the emotions and real/fake classes are presented in figure 4.4a and 4.4b respectively. In the emotion recognition task, it is noticeable that the model mostly confuses recognising correctly the emotions: disgust, contempt and angry. However, in every case, the model achieved a higher score of the true label in the confusion matrix. In the real vs fake confusion matrix, the model is struggling more to determine if a given emotion is real or fake. A possible reason for that is the lack of enough data and diversity to training the model.



(a) Confusion matrix of the 6 emotion classes.



(b) Confusion matrix of real vs fake classes.

Figure 4.4: Confusion matrix for emotions and for real/fake classes of InceptionResNetV1 best results.

4.5 Temporal Learning approach results

The results from the temporal approach (Table 4.7), demonstrate the performance of various temporal deep learning models for real/fake emotion recognition based on videos of 12 different emotions. The first four models are variants of a 3D convolutional neural network (CNN) that use different optimizers for training - Adam, RMSprop, Adagrad, and SGD. The best accuracy achieved by any of these models is only 13.333%, which is quite low. However, it's worth noting that the 3D CNN architecture is quite simple, and these models did not use any pre-trained weights. The next model is also a 3D CNN, but it includes regularization to prevent overfitting during training. This model achieves an accuracy of 15.121%, which is an improvement over the previous models. The LSTM model also achieves an accuracy of 13.333%, which is similar to the best result from the simpler 3D CNN models. However, LSTM is a more complex architecture that is designed to capture temporal dependencies in sequential data, such as videos.

Model	**Optimizer**	**Weights**	**Loss**	**Accuracy**
3D-CNN	Adam	-	0.414	8.333
3D-CNN	RMSprop	-	0.105	9.167
3D-CNN	Adagrad	-	0.105	10.833
3D-CNN	SGD	-	0.110	13.333
3D-CNN	SGD + Regularization	-	0.107	15.121
LSTM	SGD	-	0.303	13.333
X3D[15]	SGD	Kinetics 400	0.537	17.130
3D-RESNET (slow_r50)[16]	SGD	Kinetics 400	0.543	23.333

Table 4.7: Temporal approach results.

The X3D model uses the pre-trained weights from the Kinetics 400 dataset. Kinetics 400 is a large dataset of action recognition videos, and using pre-trained weights from this dataset can improve the performance of models trained on related tasks. The X3D model achieves an accuracy of 17.130%, which is better than any of the previous models. Finally, the 3D-ResNet (slow_r50) model also uses pre-trained Kinetics 400 weights and achieves the best performance among all the models, with an accuracy of 23.333%. 3D-ResNet is a deeper architecture that uses residual connections to enable the training of much deeper networks. The slow_r50 variant of 3D-ResNet is a specific configuration that balances performance and computational efficiency.

In summary, the results show that using pre-trained weights and more complex architectures can significantly improve the performance of models for real/fake emotion recognition from videos. The 3D-ResNet (slow_r50) model achieves the best performance among the tested models, but further improvements could be made with more advanced architectures and larger datasets.

Chapter 5

Discussion

Real vs fake emotion recognition is a complex and challenging task that requires the analysis of ultra-fine image details to determine the authenticity of emotion. While selecting a suitable baseline model is important, the size of the dataset and the pre-trained weights that share common features with the task data are also crucial factors. Despite the availability of state-of-the-art model architectures, they often underperform in this task due to the peculiarities of the data. Many of the exceptional architectures available today are trained in generic benchmark datasets with a variety of labels and not in datasets that are specifically designed for real and fake emotions. To overcome this challenge, it may be necessary to modify these architectures by retraining them from scratch using larger-scale, emotion-specific datasets. This is demonstrated when using VGGFace2 dataset weights instead of ImageNet. When using the formal one the performance is immediately increased than using models pre-trained on the latter one. Hence, one way to improve the performance of such models is to use the pre-trained weights of similar datasets and fine-tune them.

Furthermore, we examine how Multi-task learning could improve the performance of models in recognizing real from fake emotions. Through the experiments, is shown that it can help improve the overall performance, however not significantly. In particular, the overall accuracy of the InceptionResNetV1 pre-trained on the VGGFace2 dataset increased from 29.062% to 30.150% in the MTL approach. Nevertheless, multi-task learning is a powerful approach that can combine several tasks simultaneously, providing additional information and better insights into the given problem. For instance, in real vs fake emotion recognition, this approach can provide a better overview of how the model is performing in recognizing emotions and their authenticity.

When it comes to multi-task learning, the choice of the loss function is a crucial factor that can greatly impact the overall performance of the model. Each task in the multi-task approach requires a specific loss function that can accurately capture the objective of that particular task. Additionally, the definition of the total loss function calculation is also important, as it determines how the losses of all the tasks are combined to form a single overall loss function. In recent experiments on real vs fake emotion recognition, the Multi-Focal loss function3.2.2 was combined with other loss functions to determine the best combination for this task. The results showed that the lowest total loss and overall accuracy score was achieved when two Multi-Focal loss functions were used for each task, combined with the precision total loss3.3.

These findings suggest that while the Multi-Focal loss function is a promising choice for this task, there are still other factors that need to be considered to improve the overall performance of the model. For example, it may be necessary to further tune the hyper-parameters of the

model or to use a larger dataset to train the model. Additionally, the complexity of the task itself may require more sophisticated techniques and architectures that can capture the subtle nuances of real and fake emotions. Finally, the choice of the loss function is an important factor in multi-task learning approaches, but it is not the only factor that determines the overall performance of the model. Further research is needed to determine the optimal combination of techniques and architectures that can achieve high accuracy in real vs fake emotion recognition tasks.

In the field of emotion recognition, understanding the temporal aspect of emotions is crucial for accurately distinguishing between genuine and fake expressions. Unlike static images, which capture a single moment in time, video sequences offer a more comprehensive representation of emotions by capturing the dynamics and evolution of expressions over time. Therefore, it is essential to consider the temporal aspect of emotions when developing models for real vs fake emotion recognition tasks. One important question in this regard is which temporal architecture is best suited for this task, given the unique challenges of capturing and analyzing temporal patterns in video data.

In this context, some of the most popular temporal architectures are being examined to recognise the differences between real and fake emotions. The experiments of the temporal learning approach include comparisons with optimizers, such as Adam, Stochastic Gradient Descent (SGD), RMSprop, and Adagrad [45]. Also, temporal architectures of 3D-CNN models are utilized and trained combined with those optimizers on the provided dataset. When simple models were used without pre-trained weights the accuracy was below 15% which is insufficient for the proposed task. However, when the model complexity increased and pre-trained weights have utilized the accuracy climbed to 23.333%. That proves that larger pre-trained models could be fine-tuned in such a dataset for recognizing the authenticity of a given emotion. Those experiments are summarized in table 4.7. Several methods could lead to higher results. First of all, most of the temporal models are data sensitive and require larger datasets to fine-tune. This could immediately lead to much better results, preventing overfitting. Previous approaches use a tremendous amount of data that also requires huge computation power.

It is difficult to answer the question of which is the best-suited temporal architecture for this task. From the results, 3D-RESNET architecture or similar larger SlowFast networks[16] seems to be a promising choice for video recognition. Nevertheless, the rapid pace development of this sector requires continuous research for newly developed model architectures.

Chapter 6

Conclusion

Emotions are an integral part of human communication, and the ability to accurately perceive them plays a vital role in social interactions. With the rise of digital media, including videos, the question of whether emotions expressed through these mediums are genuine or not has become increasingly relevant. One of the main challenges in distinguishing real from fake emotions in videos is that there are no clear-cut rules for what constitutes genuine emotion. Emotions can be expressed in many different ways, and their interpretation is influenced by various factors such as cultural background, individual differences, and context. This complexity makes it difficult to establish a definitive set of criteria for differentiating between real and fake emotions in videos.

My master thesis aims to enhance the accuracy of single-task learning models in emotion recognition from videos through the use of transfer learning and fine-tuning techniques. In the STL approach, the best model achieved 4.655 and 29.062 loss and accuracy respectively. Additionally, a multi-task learning pipeline was proposed to leverage shared information in the dataset. Experiments were conducted to identify appropriate loss functions for each task, and a combined precision loss was defined to calculate the overall loss, which was found to be effective in reducing errors and improving overall performance, surpassed the STL best results with with 30.150% accuracy and 1.265 loss. Finally, the investigation of the temporal aspect of recognizing real or fake emotions in videos revealed that considering subtle information, such as the movement of facial landmarks throughout the video, can lead to better performance in this task. In the temporal context, the best accuracy was 23.333% and the loss of 0.543, with the 3D-ResNet architecture, pre-trained on Kinetics 400 dataset. The results indicated that this approach has significant potential to improve the accuracy of emotion recognition models in this field.

To further improve the performance of emotion recognition models in videos, future work can explore several avenues. One such direction is to examine transfer learning of temporal models, as this could help improve the ability of models to recognize subtle changes in facial expressions over time. Additionally, recent advancements in temporal model architectures that are not data-sensitive could be leveraged to further enhance the accuracy of these models. Another area for investigation is the use of facial landmarks, which has been found to be a promising approach for improving emotion recognition accuracy. By incorporating the movement of facial landmarks over time, models can better capture the dynamics of facial expressions and thus achieve higher accuracy.

In addition, a multi-task learning temporal approach could be an interesting path for further exploration. While similar approaches exist in Natural Language Processing, the potential benefits of this method have yet to be fully realized in the Computer Vision domain. By training models on multiple tasks simultaneously, models can learn to better identify and utilize shared

information across tasks, potentially leading to even higher accuracy and better generalization. Overall, the continued exploration and improvement of emotion recognition models in the video have significant potential to impact a variety of fields, including affective computing, human-computer interaction, and artificial intelligence.

In conclusion, the difficulty of distinguishing real from fake emotions in videos is a complex and challenging issue. While progress has been made in developing methods for identifying genuine expressions, there is still much to learn about how emotions are expressed and perceived through digital media.

Bibliography

[1] A Oyesoji Aremu and G Ayobami Lawal. A path model investigating the influence of some personal–psychological factors on the career aspirations of police trainees: a perspective from oyo state, nigeria. *Police Practice and Research: An International Journal*, 10(3):239–254, 2009.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1859–1866, 2014.

[3] V Kishore Ayyadevara and V Kishore Ayyadevara. Gradient boosting machine. *Pro machine learning algorithms: A hands-on approach to implementing algorithms in python and R*, pages 117–134, 2018.

[4] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436, 2016.

[5] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

[6] Ray L Birdwhistell. Communication without words. *Ekistics*, pages 439–444, 1968.

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[8] Allison Bruce, Illah Nourbakhsh, and Reid Simmons. The role of expressiveness and attention in human-robot interaction. In *Proceedings 2002 IEEE international conference on robotics and automation (Cat. No. 02CH37292)*, volume 4, pages 4138–4142. IEEE, 2002.

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[10] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] Jason J Dahling and Luis A Perez. Older worker, different actor? linking age and emotional labor strategies. *Personality and Individual Differences*, 48(5):574–578, 2010.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[17] James J Gross and Robert W Levenson. Eliciting emotions using films. *Cognition and Emotion*, 9(1):87–108, 1995.

[18] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 732–741, 2021.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Andreas Hennenlotter, Christian Dresel, Florian Castrop, Andres O Ceballos-Baumann, Afra M Wohlschläger, and Bernhard Haslinger. The link between facial feedback and neural activity within central circuitries of emotion—new insights from botulinum toxin–induced denervation of frown muscles. *Cerebral Cortex*, 19(3):537–542, 2009.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[24] Xuan-Phung Huynh and Yong-Guk Kim. Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3065–3072, 2017.

[25] Derek M Isaacowitz. Mood regulation in real time: Age differences in the role of looking. *Current Directions in Psychological Science*, 21(4):237–242, 2012.

[26] Shan Jia, Shuo Wang, Chuanbo Hu, Paula J Webster, and Xin Li. Detection of genuine and posed facial expressions of emotion: databases and methods. *Frontiers in Psychology*, 11:580287, 2021.

[27] Lucy Johnston, Lynden Miles, and C Neil Macrae. Why are you smiling at me? social functions of enjoyment and non-enjoyment smiles. *British Journal of Social Psychology*, 49(1):107–127, 2010.

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[29] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[30] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[31] Kaustubh Kulkarni, Ciprian Adrian Corneanu, Ikechukwu Ofodile, Sergio Escalera, Xavier Baró, Sylwia Hyniewska, Jüri Allik, and Gholamreza Anbarjafari. Automatic recognition of facial displays of unfelt emotions. *IEEE Transactions on Affective Computing*, 12(2):377–390, 2021.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[33] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, 56(4):754–772, 2006.

[34] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[35] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[37] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 15–21, 2007.

[38] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[39] Ravi Teja Mullapudi, William R Mark, Noam Shazeer, and Kayvon Fatahalian. Hydranets: Specialized dynamic architectures for efficient inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2018.

[40] Magalie Ochs, Radosław Niewiadomski, Catherine Pelachaud, and David Sadek. Intelligent expressions of emotions. In *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1*, pages 707–714. Springer, 2005.

[41] Maureen O'sullivan, Mark G Frank, Carolyn M Hurley, and Jaspreet Tiwana. Police lie detection accuracy: the effect of lie scenario. *Law and human behavior*, 33(6):530, 2009.

[42] Stephen Porter and Leanne Ten Brinke. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological science*, 19(5):508–514, 2008.

[43] Rebecca E Ready, Gennarina D Santorelli, and Molly A Mather. Judgment and classification of emotion terms by older and younger adults. *Aging & mental health*, 21(7):684–692, 2017.

[44] Johannes Reschke and Armin Sehr. Face recognition with machine learning in opencv_ fusion of the results with the localization data of an acoustic camera for speaker identification. *arXiv preprint arXiv:1707.00835*, 2017.

[45] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[46] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[47] Aakash Saroop, Pathik Ghugare, Sashank Mathamsetty, and Vaibhav Vasani. Facial emotion recognition: A multi-task approach using deep learning. *arXiv preprint arXiv:2110.15028*, 2021.

[48] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021.

[49] Frerk Saxen, Philipp Werner, and Ayoub Al-Hamadi. Real vs. fake emotion challenge: Learning to rank authenticity from facial activity descriptors. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3073–3078, 2017.

[50] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[51] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

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

[53] Newton Spolaôr, Everton Alvares Cherman, Maria Carolina Monard, and Huei Diana Lee. Relieff for multi-label feature selection. In *2013 Brazilian Conference on Intelligent Systems*, pages 6–11. IEEE, 2013.

[54] George Stockman and Linda G Shapiro. *Computer vision*. Prentice Hall PTR, 2001.

[55] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[56] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[57] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[58] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[59] Jun Wan, Sergio Escalera, Gholamreza Anbarjafari, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Juri Allik, Jelena Gorbova, Chi Lin, et al. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3189–3197, 2017.

[60] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[61] Jiannan Yang, Tiantian Qian, Fan Zhang, and Samee U Khan. Real-time facial expression recognition based on edge computing. *IEEE Access*, 9:76178–76190, 2021.

[62] Zhang Zhifei Song Yang and Qi Hairong. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[63] Lutfiah Zahara, Purnawarman Musa, Eri Prasetyo Wibowo, Irwan Karim, and Saiful Bahri Musa. The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi. In *2020 Fifth International Conference on Informatics and Computing (ICIC)*, pages 1–9, 2020.