Islamic University of Technology (IUT)

Human Facial Emotion Recognition Using Multi-head Cross Attention Network and Attention Consistency

Authors

Md. Shakib Ur Rahman, 180041211
Dilir Daiyan Rafin, 180041224
Kazi Sajid Hasan, 180041226

Supervised by

Dr. Md. Hasanul Kabir
Professor, Department of CSE

Co-supervised by

Shahriar Ivan
Lecturer, Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of BSc. in CSE.

Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
A Subsidiary organ of the Organization of Islamic Cooperation (OIC)
Academic Year: 2021-2022
May, 2023

Declaration of Authorship

This is to certify that the work presented in this thesis, titled, ”**Human Facial Emotion Recognition Using Multi-head Cross Attention Network and Attention Consistency**” is the outcome of the analysis and experiments carried out by Md. Shakib Ur Rahman, Dilir Daiyan Rafin and Kazi Sajid Hasan under the supervision of Dr. Md. Hasanul Kabir, Professor of Department of Computer Science and Engineering and Shahriar Ivan, Lecturer of Department of Computer Science and Engineering, Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Authors:

____

Md. Shakib Ur Rahman, Student ID: 180041211

____

Dilir Daiyan Rafin, Student ID: 180041224

____

Kazi Sajid Hasan, Student ID: 180041226

Supervisor:

____

Dr. Md. Hasanul Kabir
Professor
Department of Computer Science and Engineering
Islamic University of Technology

Co-supervisor:

____

Shahriar Ivan
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology

Acknowledgement

First of all we would like to express our heartfelt gratitude to Almighty Allah whose blessings were with us to the completion of this thesis.

We are indebted to Dr. Md. Hasanul Kabir, Professor, Department of Computer Science and Engineering, Islamic University of Technology for providing us with insightful knowledge and guiding us at every stage of our journey. His motivation and suggestions for this thesis have been very valuable.

We are also very much grateful to Shahriar Ivan, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology for his guidance in directing us towards our goal and helping us every time when we are in need.

We want to express our deepest appreciation to the thesis committee members for their thoughtful observations, helpful criticisms, and intelligent ideas that have substantially raised the caliber of this thesis.

Finally, we would like to express our heartiest appreciation towards our family and friends for their continuous support, motivation, suggestions and help, without which we could not have completed this work.

Abstract

In this report we discuss some methodologies related to facial expression recognition(FER). Facial expressions can be recognized with the help of collective representation of multiple facial regions and proper decoding of the high-order interactions between the local features is very necessary to recognize a particular expression efficiently. However, if noise or inconsistency is present, the FER task becomes error-prone. Because of the noise involvement in the samples, the performance of a model degrades. So, it becomes compulsory to cope with the inconsistency at first. Thus we present a model which will try to recognize facial expressions with the ability to focus on multiple regions and tackle the noise involvement or annotation ambiguity by suppressing it's effect during the training.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In human correspondence, facial expressions are immediate and critical social messages[17, 19]. They communicate essential nonverbal emotional clues in interpersonal relationships when used in conjunction with other signals. Vision-based facial expression recognition has now evolved into a useful sentiment analysis tool having a wide range of practical applications. In retail, for example, an individual's appearance information is used to determine whether a human sales assistant is required[22]. Additionally, by consistently observing facial expressions, therapists are able to help patients' conditions and treatment regimens[9]. E-learning, social robots and facial expression manipulation are some other notable crucial application areas. Facial expression recognition (FER) is a technological advancement that uses computers to identify and differentiate between facial expressions. A number of large-scale facial expression datasets were developed over the years as this investigation area has grown.

In this report, we have proposed a method to classify in-the-wild facial expression images by utilizing concepts and ideas from mainly [28] and [33] with hopes of attaining better performance and better deal with ambiguity in labelling.

1.2 Problem Formulation

Facial expression recognition (FER) makes use of computers to detect facial expressions automatically. Different facial expressions share inherently similar underlying facial appearance and their differences can be difficult to discern. One of the distinctive features of FER is that there's a very delicate contention between capturing the subtle face regions accurately and achieving a collective overall picture. A model which only consists of a single attention head has trouble focusing on multiple areas of the face at once. Therefore, cross-attention network consisting of multiple attention heads can be utilized which can deploy a series of attention heads to focus on multiple facial regions simultaneously. In this report, we discuss a model[28] whose module implements each spatial and channel attentions, that permits for capturing higher-order interactions among local facial features. It also consists of an Attention Fusion Network (AFN) that ensures attentions are drawn to different facial regions before all the attention maps are combined into a single comprehensive attention map.

The performance of FER relies massively upon high-quality annotated data. Gathering the high-quality and large-scale dataset with clear unambiguous annotations is very hard. Moreover, facial images show inter-class similarity as well as annotation ambiguity which leads to confusion, even it becomes hard for a human to identify the expression. Furthermore, deep learning models have the tendency to memorize the large-scale samples which leads to over-fitting. So, we will also introduce a framework to address this issue utilizing the concept of attention consistency[33]. We demonstrate a model which can cope with noisy samples without memorizing them.

1.3 Research Challenges

We have experienced a variety of difficulties in the FER field. The issue of noisy labels is one. Due to the ambiguity of facial emotions, which is made worse by low-quality photos, label-noise poses a significant problem for facial expression recognition in the real world.[30]. The two main categories of current research are changing the principal loss function and choosing clean samples for training. The initial strategy relies on memorization, while deep neural networks attempt to fit the pure samples.[1]. Additionally, other works concentrated on sample selection or reweighting to reduce dataset uncertainty. Using clean data Mentor net was trained in this study[27], and after training, this model was utilized to direct the Student net by reweighting the samples. Some authors suggested regularizing attention scores as well as completely reweighting the dataset. But memorizing causes over-fitting in the majority of models, which is a problem. The second approach focuses on determining the noise transition matrix or calculating the loss function. The link between the noisy labels and the actual data is modelled using the noise transition matrix. On the other hand, To suppress the noisy labels, a generalized cross-entropy loss function was proposed by Thulasidasan et al. [24] and Zhang et al. [34]. These models, however, are likewise unable to manage a huge number of classes.

Since different expressions can be similar to one another, it is harder to distinguish them. The majority of publications use single-head attentions to capture the face regions, however these models suffer from over-lapping and are unable to gain enough information about critical facial areas. Class separability should be applied appropriately to discover the distinctions. However, [28] proposed a multi-head cross attention module to address the mentioned issue.

Chapter 2

Literature Review

2.1 Facial Expression Recognition

One of the most amazing and challenging study tasks in social correspondence is recognizing human emotion from images. The performance of deep learning (DL) based emotion recognition is superior to that of traditional image processing techniques.[12] applied a robust CNN to picture identification and used the deep learning open-source library ”Keras” developed and provided by Google for facial emotion detection. For the purpose of recognizing human emotions, a face-sensitive convolutional neural network (FS-CNN) was proposed by [20] to recognize faces in large-scale images, after which face milestones are investigated to forecast appearances for feeling acknowledgment. Also PRATIT, a model for facial expression detection that uses particular image preprocessing processes and a Convolutional Neural Network (CNN) model, has been proposed by [16].

For human computer interaction to be successful in advancing applications for artificial intelligence and humanoid robots, real-time facial recognition applications must have the option to be done at a rapid and accurate pace. Real-time video data was used to identify faces with deep learning techniques in [4] to identify the faces’ expressions of happiness, surprise, surprise, sadness, and neutral emotion.

[27] suggested an approach called Self-Cure Network (SCN) to suppress the vulnerabilities for large-scale facial expression recognition in order to address noisy-label or ambiguity concerns.

[32] stated that uncertainty is a relative idea. Motivated by this idea, they presented Relative Uncertainty Learning, a novel technique for aiding deep learning models for the purpose of learning uncertainty for each sample. They carefully created a second branch to show the input image uncertainty and used it to reduce noise during training.

2.2 Attention Mechanism

In visual perception, attention mechanism plays an important role [3, 18]. Particularly, attention enables individuals to efficiently search for more significant information in a complex setting. There have been some recent attempts to replace CNNs and Recurrent Neural Networks (RNN) with monotonous attention modules that have produced impressive results. For instance, [25] suggested a straightforward network architecture known as Transformer that completely avoids recurrence and convolutions in favor of a multi-head attention mechanism. Based on Transformer [6] established a pure attention mechanism whose effectiveness outperformed earlier state-of-the-art techniques. Convolutional Block Consideration Module (CBAM) has been suggested by [29] as a method for gaining rich attention features by sequentially integrating channel and spatial attention.

2.3 Discriminative Loss

Recent research demonstrates how the discriminative loss function can be significantly tailored to the FER problem. A DDA loss was proposed in and the benefits of center loss and softmax loss were combined in [7]. Similarly, to boost the inter-class distance for various categories, a cosine metric based on center loss was introduced in [2]. Additionally, an attentive center loss that promotes understanding the connections between each class center and the center loss was put forth in [8]. All these loss functions introduce additional parameters and computations. The affinity loss is proposed in [28] is more straightforward, and the inter-class distance is widened by using the internal relationship between the class centers.

2.4 Label Noise

Since the recognition inconsistency in the lab-gathered FER datasets is significantly high, additional initiatives have been taken in recent years to address the FER problem for in-the-wild (ITW) images, which have a lot of label noise. To further mine the underlying truth, [31] first took annotation inconsistencies into account and gave each sample more than one label. Wang et al. It was proposed by [27] to relabel the photos in order to suppress the uncertain ones and learn an important weight for each sample. [21] trained multi-branch models, leaving out one class for each branch, to identify the latent truth under label noise. [32] proposed to learn the uncertainty of different facial images by comparison and then suppress the uncertain images. They can be mainly categorized into two classes, sample selection [27, 32] or label ensembling [31, 21]. Sample selection is a technique for choosing clean samples in order to filter out noise among the dataset. On the other hand, label ensembling is a technique that utilizes crowdsourcing methods to achieve better performance. To sum up, it is required to know either noise rate in order to channel out noisy samples or incur additional calculation overhead for the classifications tasks that deal with a large amount of classes. However, the aforementioned methods cannot generalize the classification tasks properly. Thus, Zhang et. al. [33] proposed erasing-attention consistency(EAC) method which prevents the model from over-fitting without knowing the underlying noise rate present in the dataset or introducing extra calculation and is better in generalizing classification tasks too.

2.5 Attention Consistency

The idea of attention consistency is initially proposed by [10] in order to improve multi-label image classification and visual perceptual plausibility by taking into account visual attention consistency under spatial transforms. It is assumed that the attention maps that were learnt by the model would undergo similar transformation as those for the input images. This allows a comparison between the attention maps to determine the degree of ambiguity.

Chapter 3

Methodology

Our proposed model utilizes a multi-headed cross attention network for classification and an attention consistency loss to reduce the impact of noisy labels on the training. The cross attention network was inspired by the work of [28] and the concept of attention consistency was obtained from [10]. Below we show the architecture of our proposed model. After that, we explain the workings of the two parts of our model as mentioned earlier.

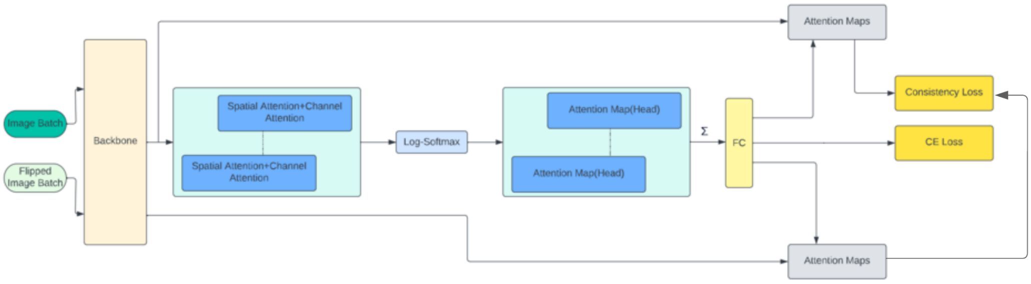

Figure 3.1: **Our proposed model.** The cross-attention network is used to obtain the cross-entropy or classification loss and the attention consistency loss is calculated from the extracted features and the weights from the last FC layer.

3.1 Cross Attention Network

The cross-attention network, as in [28] is broken down into three parts to teach high-quality attentive features: Multi-head cross Attention Network (MAN), Attention Fusion Network (AFN), and Feature Clustering Network (FCN). First, from a batch of face expression images, the input images are erased randomly. Then these images are flipped and two batches of images are obtained, one containing the original images I and the other containing the corresponding flipped images $I^{'}$. Both of these batches of images are passed through the subsequent FCN, MAN and AFN modules. The FCN outputs a basic feature embedding with class discrimination capabilities from the images. After that, several sectional facial expression regions are captured using the MAN. The AFN then normalizes the attention maps and trains these maps to focus on various areas to avoid overlapping. Lastly, the AFN combines these aforementioned attention maps into one to predict the expression class of input images.

MAN contains two types of attention units. Spatial attention unit and channel attention unit. Spatial attention units contain convolutional kernels to extract image features in various scale and channel attention units provide encoder-decoder mechanism for the spatial features. The process has been shown in figure 3.2.

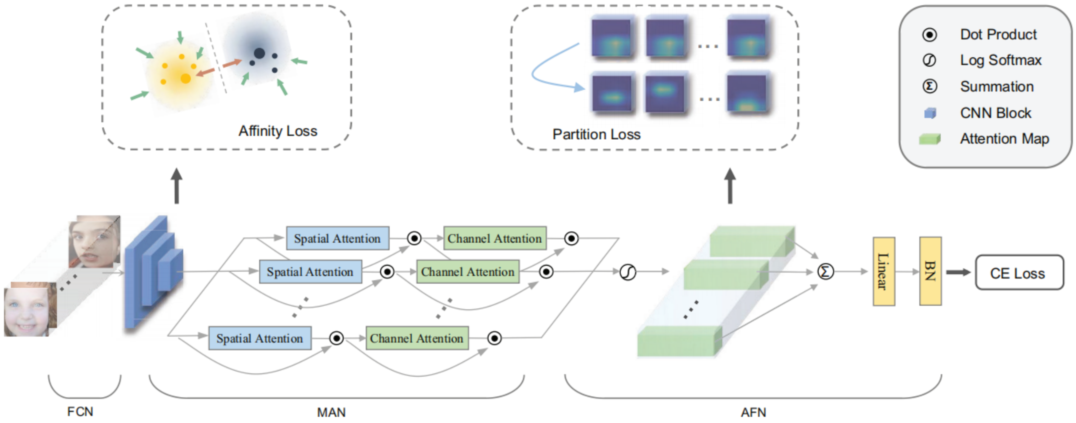

Figure 3.2: **Overview of the Cross-Attention Network**. The approach is divided into three sub-networks. The essential features are extracted and grouped first by FCN, which utilizes a kind of loss function called affinity loss that decreases the intra-class distance and also increases the inter-class distances. After that, there is MAN which deploys multiple attention heads to focus on multiple interesting face regions. Finally, there is the AFN which which utilizes partition loss to make the heads focus on different regions and leads to the classification.

3.1.1 Feature Clustering Network

Though facial expressions can be different but these expressions can share some similarities. So, the differences may become subtle and proper discrimination between the underlying features become necessary. To maximize class margin, affinity loss[28] is used which is a kind of discriminative loss. It tries to increase the inter-class distance and reduce the intra-class distance. It does so by picking class centres for each class through random sampling on an n-dimensional gaussian distribution, where n represents the dimension of class centres. This allows the model to learn a good differentiation between classes.

$$\mathcal{L}_{af} = \frac{\Sigma_{i=1}^{Y} ||f_i - c_{yi}||_2^2}{\sigma_c^2} \tag{3.1}$$

In equation 3.1, Y is the dimension of the label space, f_i represents the feature vector obtained from the backbone for the i-th input vector, c_{y_i} represents the class centre for the label y_i and σ_c denotes the standard deviation among class centres.

3.1.2 Multi-head Cross Attention Network

The MAN module consists of some parallel attention heads that work independently of each other. Each attention head consists of a spatial attention unit followed by a channel attention unit. The feature vectors from the backbone network are taken as input into each head. The outputs from the spatial attention unit are taken as input into the channel attention unit which then generates the final attention map for a particular head. The structure of an attention head has been shown in figure 3.3.

From the figure 3.3, we can see that the spatial attention network consists of 4 convolutional kernels with different size and one activation function. These kernels are used to capture image features at different scales. On the other hand, the channel attention units contain a pooling layer and two linear layers with one activation function.

Let h denote the number of cross-attention heads, $A_s = \{A_{s_1}, A_{s_2}, ..., A_{s_h}\}$ and $A_c = \{A_{c_1}, A_{c_2}, ..., A_{c_h}\}$ denote the spatial attention units and channel attention units of each attention head. Also, let $M_s = \{M_{s_1}, M_{s_2}, ..., M_{s_h}\}$ denote the spatial attention maps generated by each spatial attention unit and $M_c = \{M_{c_1}, M_{c_2}, ..., M_{c_h}\}$ denote that generated by each channel attention unit. Then the output from the i-th spatial attention unit is given by-

$$\underset{i \in \{1,..,h\}}{M_{s_i}} = f \times A_{s_i}(w_s, f) \tag{3.2}$$

where w_s are the network parameters of $A_{s_i} and.$

Similarly, the output of the i-th channel attention unit is given by-

$$\underset{i\in\{1,..,h\}}{M_{c_i}} = M_{s_i} \times A_{s_i}(w_c, M_{s_i}) \tag{3.3}$$

where w_c are the network parameters of A_{c_i}.

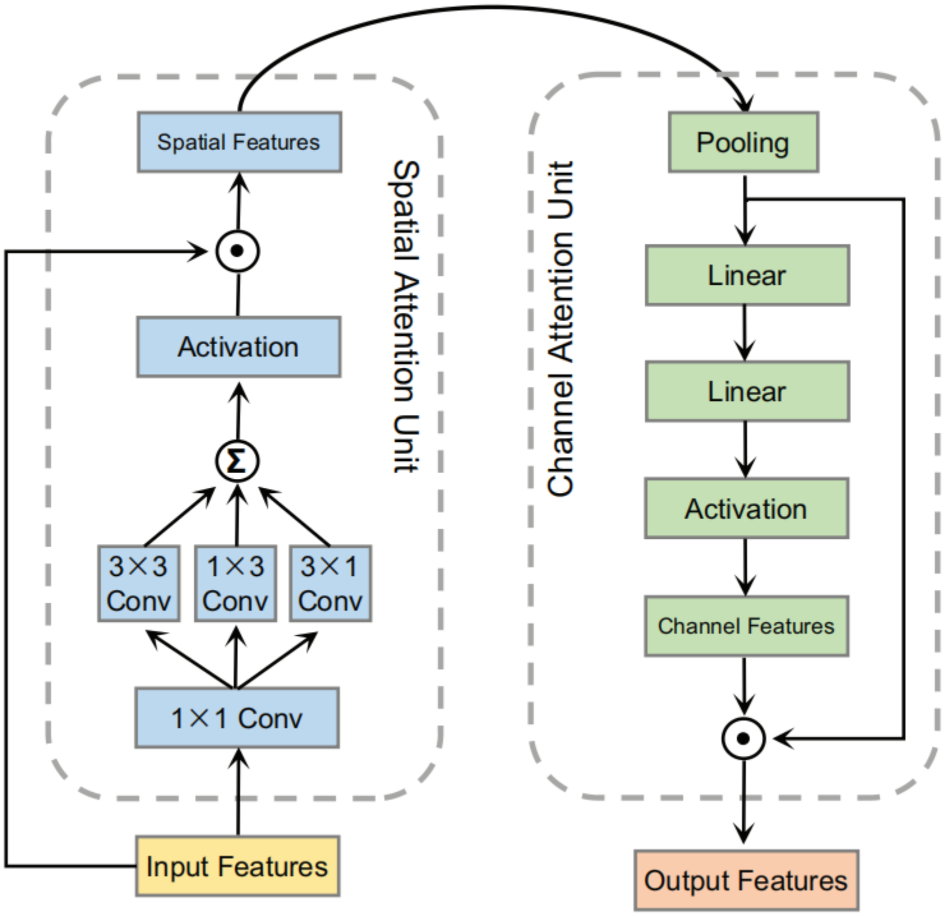

Figure 3.3: **Structure of an attention head.** It comprises of a spatial attention unit followed by a channel attention unit.

3.1.3 Attention Fusion Network

Though cross-attention heads can capture features in a better way but still may suffer from over-lapping. Over-lapping happens when multiple attention maps focus on same image regions. To handle this, a log-softmax function is first used to scale the attention maps generated by the attention heads. Next, a partition loss is used which penalizes the similar attention maps from the independent heads and forces the attention heads to focus on different regions to avoid overlapping. Finally, the normalized attention maps are merged into a single attention map and it is directed to the final fully-connected layer for the computing class prediction.

Let, v denote the output vectors of the i-th cross attention head H_i, v_j denote the j-th vector of H_i, then $v^{'}$, the output from applying the log-softmax is computed as:

$$\underset{j \in \{1,..,h\}}{v_j^{'}} = \log(\frac{e^{v_j}}{\Sigma_{k=1}^{h} e^{v_k}}) \tag{3.4}$$

The difference between the attention maps generated by the cross-attention heads is maximized by a partition loss. In particular, h, the number of cross-attention heads has been used to determine the ideal number of different unique regions to focus on. Additionally, the MAN with more cross-attention heads may be able to focus on more subtle areas which will lead to extraction of higher level facial features. Then, the partition loss can be written as follows:

$$\mathcal{L}_{pt} = \frac{1}{NC} \times \Sigma_{i=1}^{N} \Sigma_{j=1}^{C} \log(1 + \frac{h}{\sigma_{ij}^{2}}) \tag{3.5}$$

where C is the channel size of the attention maps, $\sigma_{i,j}^{2}$, j denotes the variance of the j-th channel on the i-th sample.

3.2 Calculating the Flip Attention Consistency

The alignment or coherence between a neural network's attention mechanism and human attention is referred to as attention consistency. It seeks to make sure that the model pays attention to pertinent portions of the input data, much as how people pay attention to significant areas. The interpretability and robustness of the model's predictions are both enhanced by this consistency.

The facial images before and after flip have the same semantic meaning related to the facial expression[33]. The flip attention consistency refers to the idea that if input images are made to undergo some kind of a spatial transformation, then the attention maps generated by the model should also be according to the same transformation as the input images. It includes using human eye-tracking data to build a model that anticipates where people will focus their attention. Utilizing this concept, we apply the flip attention consistency loss[33] to prevent the model from learning from the noisy label samples.

3.2.1 Attention Map Generation

The attention map produced by the model tells us what the model is basing it's predictions on. It's a weighted sum of feature maps where the weights were extracted from the last convolution layer. Let the feature map obtained from the final convolution layer be denoted by $F \in \mathbb{R}^{C \times Y \times X}$. Here, C represents the number of channels, Y represents the height and X represents the width of the feature maps. Let the weights from the FC layer be denoted by $W \in \mathbb{R}^{L \times C}$, where L denotes the no. of classes. Then the attention map is calculated as follows:

$$M_i(y, x) = \Sigma_{c=1}^{C} W(i, c) F_c(y, x) \tag{3.6}$$

Here, $M_i(y, x)$ is the value of the attention at location (y, x) for class index i. We extract the weights from the last FC layer in AFN after passing through the original images I. These weights are used to generate attention maps not only for the feature maps of I but also for those of $I^{'}$. We denote the generated attention maps from I as M and those from $I^{'}$ as $M^{'}$.

3.2.2 Flip Attention Consistency Loss

Let W_{a_i} denote the a_i-th weight from the final FC layer of AFN with a_i as the label of the i-th image. Then we use the below equation to calculate the flip attention consistency loss l_c which minimizes the distance between the two attention maps M and $Flip(M^{'})$:

$$\mathcal{L}_c = \frac{1}{NLYX} \Sigma_{p=1}^{N} \Sigma_{q=1}^{L} \| M_{pq} - Flip(M^{'})_{pq} \|_2 \tag{3.7}$$

The combined loss function comprising of the affinity loss used in FCN, the partition loss used in AFN, the flip attention consistency loss and the cross-entropy loss used to classify can be written as-

$$\mathcal{L} = \lambda_1 \mathcal{L}_{af} + \lambda_2 \mathcal{L}_{pt} + \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_c \tag{3.8}$$

where λ_1, λ_2 are the weighting hyper-parameters for $\mathcal{L}_{af}$ and $\mathcal{L}_{pt}$. Here, both λ_1 and λ_2 are empirically set to 1.0 and λ_3 is set to 4.

Chapter 4

Experiments

In this section, our experimental evaluation results are described in details. The proposed method has been used in three famous FER datasets: RAF-DB, AffectNet-7 and AffectNet-8. We have gained a slightly higher accuracy in RAF-DB but slightly less accuracy in both the classes of AffectNet datasets in comparison to DAN[28].

4.1 Datasets

4.1.1 RAF-DB

RAF-DB [14] is a large-scale facial expression dataset comprising of two different subsets of images- basic and compound emotions. The basic subset contains images of seven classes- surprised, fearful, happy, sad, angry, disgusted and neutral. There are 15,339 images in the basic set with 12,271 images in the training set and 3,068 images in the test set. The compound subset contains images of 12 compound emotion classes.

4.1.2 AffectNet

AffectNet [5] is another large-scale facial expression dataset. It is the largest benchmark dataset for ITW face images in this domain and consists of seven classes- happy, sad, surprise, fear, disgust, anger, neutral, contempt. The first seven classes form AffectNet-7 which contains 287,401 images among which 283,901 are training images and 3,500 are validation images. There is also AffectNet-8 which contains the seven classes from AffectNet-7 along with the contempt class. It contains 287,651 training images and 3,999 validation images.

4.2 Implementation Details

The aligned images provided by both RAF-DB and AffectNet were used for training and validation of the model. Images were reshaped according to the backbone used during training and validation, in this case, 224 x 224. In order to avoid overfitting, some common data enhancement techniques like rotation, flip etc. were also used. Since ResNet-18[11] performed the best as the backbone, it was adopted as the backbone feature extractor for the model.

The code was written using the python language and the Pytorch framework and the model was trained for 40 epochs on a personal computer having an RTX 3070 8GB gpu for all the experimentations and results. The number of attention heads in the cross-attention network was set to 4 as it seemed to give the best performance.

For the RAF-DB dataset, Adam was used as the optimizer with a batch size of 64 and a learning rate of 0.0001. For AffectNet-7 and AffectNet-8 datasets, also Adam was used as the optimizer with a batch size of 32 and a learning rate of 0.0001. Since there is an inconsistent ratio in the data of the training and validation sets of both datasets, some sampling strategy was used to upsample the classes having less volume and downsample the ones having very high volume.

This helps to attain a balance between the classes during training time.

4.3 Ablation Studies

Ablation studies [15], also known as sensitivity analysis or feature removal analysis, are a method frequently employed in machine learning and scientific research to comprehend the value and contribution of various parts or aspects within a system. The term "ablation" describes the process of deliberately deleting or disabling certain components or factors to examine the influence on the performance or behavior of the system as a whole. To verify and understand the performance and effects of the different components of the model, ablation studies were performed on the RAF-DB[14] dataset.

4.3.1 Number of Cross Attention Heads

The number of attention heads used has an effect on the performance of the model. As we can see in Figure 4.1[28], the performance of the model seems to vary based on the number of attention heads used. Multiple attention heads seem to improve the performance over using a single one and 4 seems to be the optimal number of attention heads. Figure 4.1 holds similarly for our modified model as well.

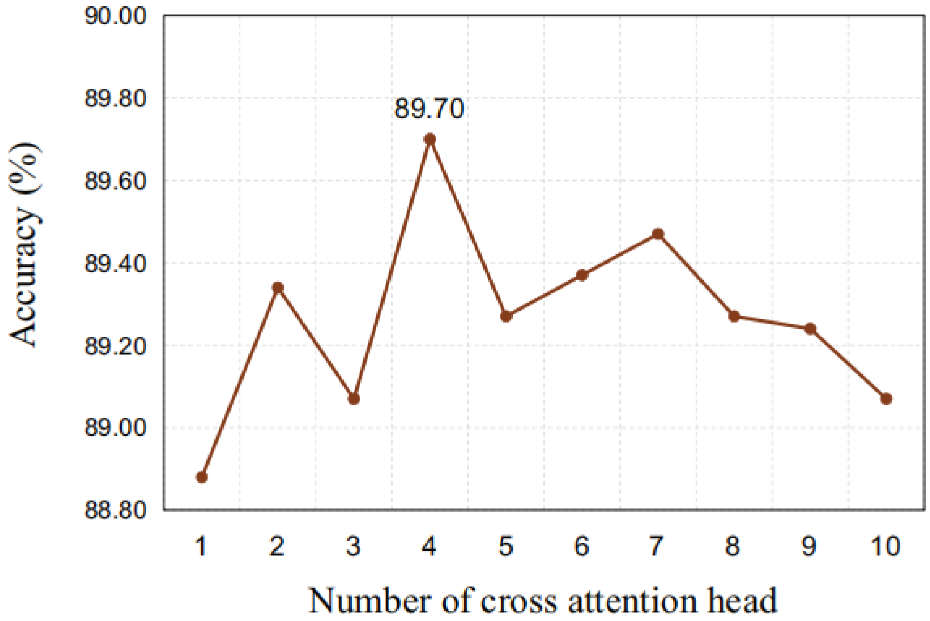

Figure 4.1: Ablation studies for different numbers of attention heads in the MAN module for RAF-DB dataset [28].

4.3.2 Effects of Different Loss Functions Used

The usefulness and effects of the two loss functions, affinity loss in FCN and the partition loss in AFN, have been tested separately in the tables 4.1 and 4.2, while the effects of the consistency loss has been shown in table 4.3 on RAF-DB with 30% noisy labels.

Methods	Accuracy(%)
-	87.88
Affinity Loss	89.07

Table 4.1: Performance comparison of the model with and without the affinity loss used in FCN.

Methods	Accuracy(%)
-	87.30
Partition Loss	89.07

Table 4.2: Performance comparison of the model with and without the partition loss used in AFN.

Methods	Accuracy(%) on 30% Noise
-	79.11
Consistency Loss	81.46

Table 4.3: Performance comparison of the model with and without the consistency loss.

4.4 Comparison With Other Models

The performance of our proposed model on RAF-DB, AffectNet-8 and AffectNet-7 datasets are given in the tables 4.4, 4.5 and 4.6 respectively along with performances of various other models on those datasets for comparison. ResNet-18 is taken as the baseline model for all 3 datasets. The model achieved an accuracy of 89.07% on RAF-DB dataset and 60.33% and 63.42% on AffectNet-8 and AffectNet-7 respectively. This shows that the proposed model is quite competitive in terms of performance even though it could not achieve state-of-the-art results.

Methods	Accuracy(%)
PSR[26]	88.98
SCN[27]	87.03
MViT[13]	88.62
RUL[32]	88.98
ResNet-18 [11]	71.67
DAN[28]	89.73
EAC[33]	89.04
Ours	89.07

Table 4.4: Performance comparison of our model with various other models on RAF-DB.

Methods	Accuracy(%)
PSR[26]	60.68
SCN[27]	60.23
MViT[13]	61.40
RUL[32]	60.66
Baseline (ResNet-18) [11]	56.84
DAN[28]	61.49
Ours	60.33

Table 4.5: Performance comparison of our model with various other models on AffectNet-8.

Methods	Accuracy(%)
PSR[26]	63.77
DDA-Loss[7]	62.34
MViT[13]	64.57
EfficientFace[35]	63.70
Baseline (ResNet-18) [11]	56.97
DAN[28]	65.20
EAC[33]	64.32
Ours	63.42

Table 4.6: Performance comparison of our model with various other models on AffectNet-7.

4.5 Comparison of our model with different backbones on RAF-DB

We have run the proposed model with various backbone feature extractors on RAF-DB dataset. The comparative performance among the different backbone models is given in 4.7. For clean samples, ResNet-18 backbone gives us the highest accuracy(89.07%) among all. Whereas Vision Transformer has acquired the lowest accuracy among them(56.71%). We couldn't identify the reason behind this poor performance of ViT.

Then we added 30% noise in the dataset and ran our model using the backbones again. This was done to test the ambiguous label handling capacity of our model. ResNet-18 [11] has achieved the highest accuracy(81.46%) here also but again Vision Transformer [6] gave us the lowest performance(48.32%).

Our Model	Clean Samples(%)	30% Noise(%)
ResNet-18[11]	89.07	81.46
ResNet-50[11]	87.39	75.57
Inception_V3[23]	86.41	73.02
ViT_B_32[6]	56.71	48.32

Table 4.7: Performance comparison of our model with different backbones on RAF-DB.

Chapter 5

Conclusion

Our work was mainly focused on the idea of multi-head cross attention network and erasing attention consistency. Multi-head cross attention network (MAN) has the ability to emphasize on multiple facial areas and it can extract crucial features more efficiently. Moreover, attention consistency encourages the model to learn more from the unambiguously labelled images and the consistency loss prevents the model from remembering noisy samples. Thus, in this paper, we have proposed an architecture by incorporating the technique of erasing attention consistency into the multi-head cross attention network to develop a good noise-handling mechanism and achieve better generalization in classification tasks. In future, we want to explore further to learn and develop better noise handling techniques and also explore the possibilities of vision transformers in this domain which seems to be getting more and more attention in recent times.

Bibliography

[1] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 7

[2] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018. 10

[3] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002. 9

[4] Emre Dandıl and Rıdvan Özdemir. Real-time facial emotion classification using deep learning. *Data Science and Applications*, 2(1):13–17, 2019. 8

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(03):34–41, 2012. 21

[6] 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. 9, 26

[7] Amir Hossein Farzaneh and Xiaojun Qi. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition Workshops, pages 406–407, 2020. 9, 25

[8] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021. 10

[9] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003. 5

[10] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019. 11, 12

[11] 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, 24, 25, 26

[12] Akriti Jaiswal, A Krishnama Raju, and Suman Deb. Facial emotion detection using deep learning. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE, 2020. 8

[13] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvt: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021. 24, 25

[14] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 20, 22

[15] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019. 22

[16] Dhara Mungra, Anjali Agrawal, Priyanka Sharma, Sudeep Tanwar, and Mohammad S Obaidat. Pratit: a cnn-based emotion recognition system using his-

togram equalization and data augmentation. *Multimedia Tools and Applications*, 79(3):2285–2307, 2020. 8

[17] Catherine Newmark. Charles darwin: the expression of the emotions in man and animals. In *Schlüsselwerke der Emotionssoziologie*, pages 111–115. Springer, 2022. 5

[18] Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000. 9

[19] Erika L Rosenberg and Paul Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, 2020. 5

[20] Yahia Said and Mohammad Barr. Human emotion recognition based on facial expressions via deep learning on high-resolution images. *Multimedia Tools and Applications*, 80(16):25241–25253, 2021. 8

[21] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021. 10

[22] Gurvinder Singh Shergill, Abdolhossein Sarrafzadeh, Olaf Diegel, and Aruna Shekar. Computerized sales assistants: the application of computer technology to measure consumer interest-a conceptual framework. 2008. 5

[23] 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. 26

[24] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*, 2019. 7

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 9

[26] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020. 24, 25

[27] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 7, 9, 10, 24, 25

[28] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 3, 5, 6, 7, 10, 12, 13, 14, 20, 22, 23, 24, 25

[29] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 9

[30] Huan Yan, Yu Gu, Xiang Zhang, Yantong Wang, Yusheng Ji, and Fuji Ren. Mitigating label-noise for facial expression recognition in the wild. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 7

[31] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 10

[32] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021. 9, 10, 24, 25

[33] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Eu-*

ropean Conference on Computer Vision, pages 418–434. Springer, 2022. 5, 6, 10, 18, 24, 25

[34] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 7

[35] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3510–3519, 2021. 25