Stiftung
Haus der Geschichte
der Bundesrepublik Deutschland

Fraunhofer
IAIS

Project Report

Facial Emotion Recognition for German Oral History Interviews

Krishna Teja Nallanukala
Msc Autonomous Systems
Hochschule Bonn Rhein-Sieg University

Supervised by

Duc Bach Ha

September 2022

Contents

List of Figures

List of Tables

1 Introduction

Emotion is an important part of daily interpersonal human interactions. Humans express their emotions in a variety of ways, including through spoken words and nonverbal cues like body language and facial expression. Recently, automatic recognition or detection of human emotions have attracted much research interest in the field of computer vision, speech processing, and multimedia computing [1]. Mainly, the video based emotion recognition because of its applications in detecting emotions in real world environments, where emotion is conveyed in a series of actions. Algorithms for emotion recognition from videos generally requires information from 2 modalities like facial expression, Audio to predict an emotion. The main project experiments using all the modalities. However, this sub project mainly focuses on dealing with only facial expression modality to investigate the possibility of detecting emotions using just single modality. Facial expressions are the most natural and universal way of conveying emotions and intentions for human beings. In the field of computer vision, Facial Expression Recognition(FER) is the process of recognizing emotions given an image/video of humans as input. In recent years, the applications of FER are distributed across various fields such as Autonomous cars, Human-Computer Interaction (HCI), education, healthcare, and psychological treatments [2]. Furthermore, the advancements in Deep Neural Network (DNN) methods, showed remarkable performance in image recognition tasks. Convolutional Neural Networks (CNN), in particular as the prominent technique that extracts deep feature representations have performed better than the traditional techniques. In this project, we aim to classify videos of interviews based on predominant emotions expressed by subjects contained in them.

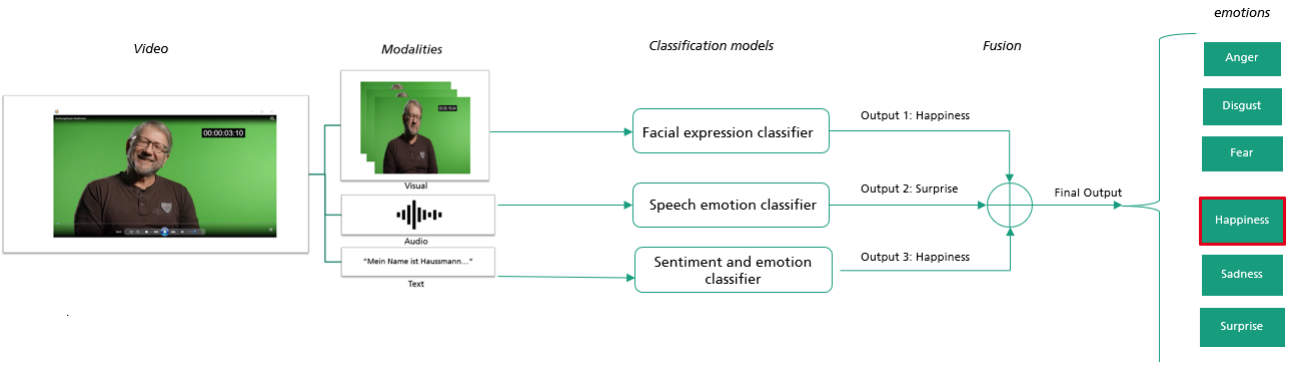


Figure 1: Illustration of process for emotion recognition from video

Generally, the feature extraction process for video based FER is mostly categorized into two types: static based methods and temporal based methods [3]. Static based feature extraction is similar to image based, where the features are extracted on every frame in the video. These features are used for building a classifier. However, the static based method doesn't consider the temporal relation between the frames, which are essential for classifying a video [3]. Additionally, the sequence based expression recognition approaches enables to detect the emotions of humans in real world environments rather than image based FER [2]. Temporal based feature extraction techniques model the temporal relation and motion information from videos. 3D convolution Networks(3DCNN), Recurrent Neural Networks (RNN), and their variants are predominantly used techniques for modeling temporal relation in videos for FER[3]. However, RNN is computationally inefficient and incapable of extracting deep features, when compared to CNN [4]. Whereas, 3d filters in 3DCNN can be used only for small video clips. Additionally, "static-based methods are superior to the others according to several winner solutions in EmotiW challenges" [3]. For the static based methods model the temporal relationship between frames, frame aggregation operation is required and should also consider the importance of frames in video [4]. Further investigation into the literature review for FER methods, shortlisted the potential set of techniques for the project. Out of the five techniques given in Figure 2, Frame Attention Networks (FAN) was chosen for this project for its high performance (accuracy) on CK+(99%) and AFEW (51.18%) datasets and its simple yet effective construction. The FAN

technique is based on the attention mechanism of the transformer network, which is used for machine translation. The FAN is designed to learn frame importance reasoning in an end-to-end fashion using the attention mechanism and enables adaptively aggregating the frame features [3].

List of Potential Methods (shortlisted from State-Of-The-Art)				
Rank (subjective)	**Paper**	**Approach category**	**Year of publication, citations**	**Performance on its evaluated datasets**
#1	Frame Attention Networks for Facial Expression Recognition in Videos	(CNN + Attention mechanism), Frame aggregation performed dynamically using attention mechanism	2019, 20	CK+ - 99% accuracy, AFEW 8.0 - 51.18% accuracy
#2	A Compact Deep Learning Model for Robust Facial Expression Recognition **(CNN + GRUs)**	(CNN+GRUs)	2018, 71	Oulu-CASIA-91.67% accuracy, GENKI-95.33%, RAF-65.52%accuracy
#3	Affect Expression Behaviour Analysis in the Wild using Spatio-**Channel Attention** and Complementary Context Information	(CNN + Attention mechanism	2020, 1	Aff-Wild2 - 73.4% accuracy, Ranks second in ABAW-2020 challenge
#3	Facial expression and attributes recognition based on multi-task learning of lightweight neural networks		2021,-	AffectNet dataset-65.74% accuracy (Ranks first)
	Two-Stream Aural-Visual Affect Analysis in the Wild		2020, 5	AFF-WILD2 - 70% accuracy, Ranks first in ABAW-2020 challenge

Figure 2: State of the art techniques for FER

2 FAN Architecture

FAN takes in a set of input faces from a video and outputs its corresponding emotion label. Therefore, frames from the HDG videos are extracted, and then fed into the d-lib based face detection and alignment modules to crop out and align faces respectively. Facial images thus obtained are used with FAN.

The network architecture described in FAN consists of two modules: feature embedding and frame attention. The feature embedding module is a ResNet-18 based Convolutional Neural Network (CNN) which converts a predefined number of face images into embeddings. The feature vectors are in turn weighted in the subsequent frame attention module to output the class label. By varying the scheme to weigh embeddings in the frame attention module, the work offers three variants of FAN: baseline, self-attention, and the combination of self and relation attention.

The baseline variant performs the task based on summing up the soft-max scores of every frame. In the case of the self-attention variant, the frames are adaptively weighted based on learned weights to form a single characteristic representation for the entire video called the "video-level global anchor" which is then used for discrimination. The relation-attention variant goes one step ahead by relating individual frame features with the global-level anchor representation to produce a single compact feature for classification. The authors of the work report slightly superior performance of the self-relation-attention variant over the other two, while using the AFEW dataset. The following Figure 3 depicts the emotion recognition pipeline used in FAN.

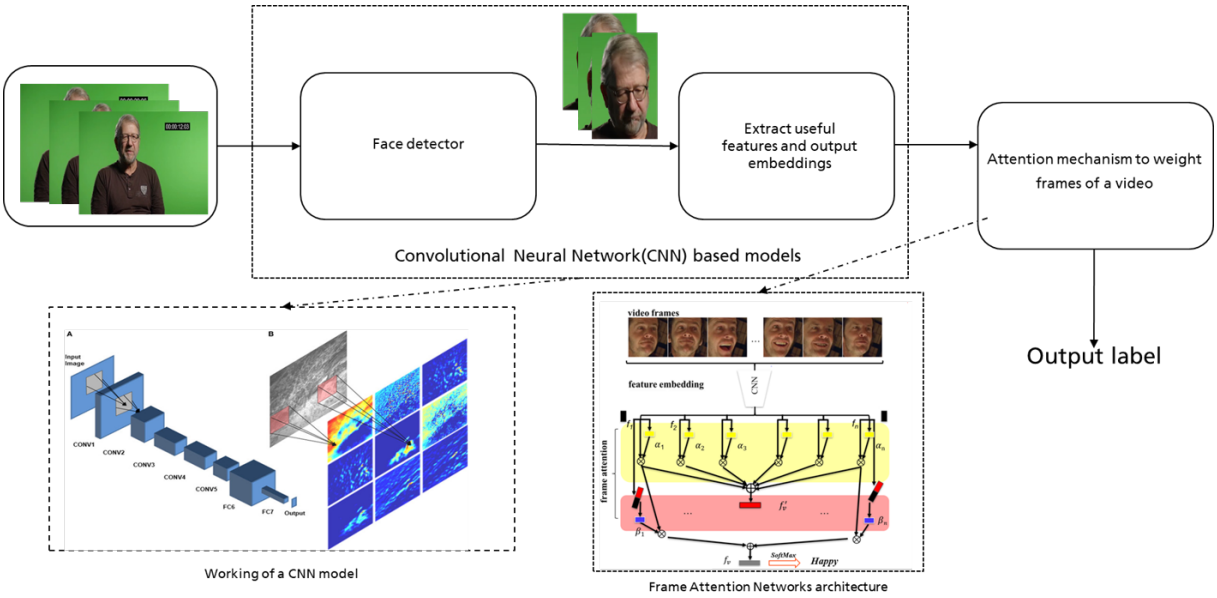
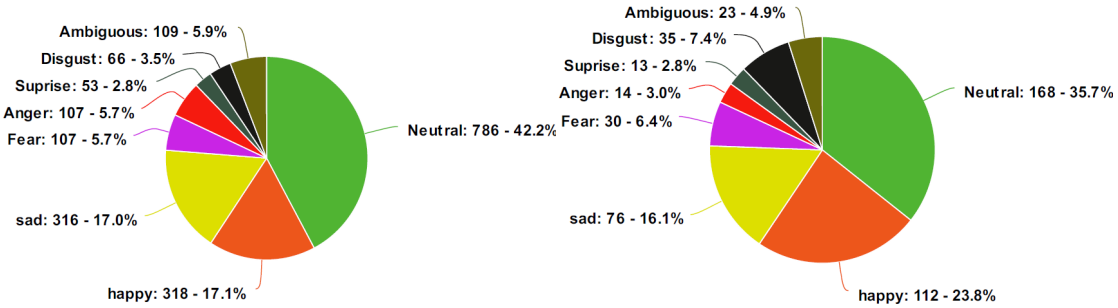


Figure 3: FAN pipeline [3]

3 Dataset

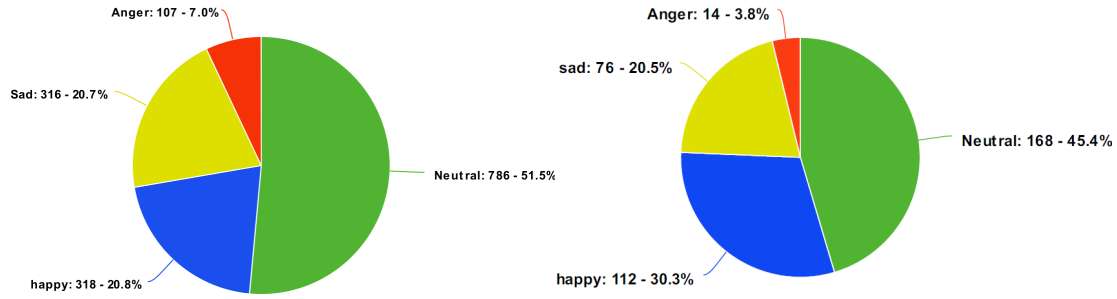
This project uses Haus der Gesichte(HDG) dataset. This dataset consists of short video clips taken from the interviews conducted by Haus der Gesichte. Videos are divided into 8 emotion classes: Happy, Anger, Surprise, Fear, Neutral, Disgust, Sad and Ambiguous. However, initial experiments only considered 4 classes: happy, sad, anger, and neutral. This particular split of this dataset is hereafter referred to

as "HDG4". Figure 4 represents the class distribution of train split and test split of the whole dataset and Figure 5 represents the class distribution of train and test split of the HDG4 dataset.



(a) Train split (b) Test split

Figure 4: Class distribution of HDG dataset



(a) Train split (b) Test split

Figure 5: Class distribution of HDG4 dataset

4 Experiments & Results

This section discusses the various experiments performed by two students worked on this project. The project is started by Ashwin Kumar Vijayananth and later continued by Krishna Teja Nallanukala.

4.1 Ashwin's experiments

This section discusses the various experiments performed to implement FAN for emotion recognition.

4.1.1 Experiment a: Reproduction of FAN authors's result

In this experiment, FAN is retrained on AFEW dataset to evaluate the claims made by authors about the performance. The authors provided the model checkpoint, which was trained on the Micrososft FER+ dataset, which he used as a prior (pre-trained model) for training on the AFEW dataset. The following Table 6 contains the results of the experiments.

Frame Attention Networks - evaluation on AFEW dataset			
Architecture	**Pretrained model**	**Model**	**Accuracy**
Baseline (without attention)	Microsoft FER+ dataset	author's (value reported in paper)	**48.82**
		locally trained	46.45
	None	locally trained	21.26
Self attention	Microsoft FER+ dataset	author's (value reported in paper)	**50.92**
		locally trained	44.88
	None	locally trained	29.65
Self and relation attention	Microsoft FER+ dataset	author's (value reported in paper)	**51.18**
		locally trained	16.53
	None	locally trained	16.53

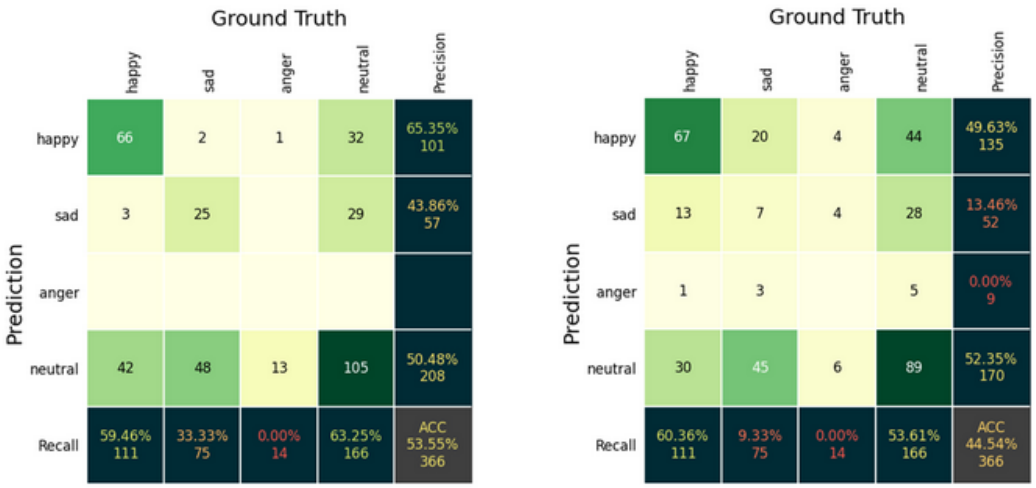
Figure 6: Comparing authors results with locally trained results

From the results, it is evident that we were able to closely reproduce the author's results for the baseline and self-attention variants on AFEW dataset. However, it was not possible with the self-relation attention variant due to some inherent problems in the source code. Based on the author's claims, the self-relation attention variant only shows a marginal improvement over the others. Owing to the need to quickly prototype FAN for the "HDG" dataset, we resorted to conducting further experiments with the baseline and self-attention variants.

4.1.2 Experiment b: Evaluate FAN on the HDG4 dataset

In this experiment, the FAN is trained on HDG4 dataset. As discussed in Section 3, initial experiments considered to use HDG4 dataset. Figure 7 illustrates the results of model variants. From the results, we can infer that the baseline variant of FAN performed better than the self-attention variant. Furthermore, both variants

exhibit poor performance in the case of anger class. This is due to the imbalance in the dataset and Section 4.1.3 analyses the imbalance problem by performing various experiments.



(a) Baseline variant (b) self attention variant

Figure 7: Confusion matrices of model variants on HDG4 dataset

4.1.3 Experiment c: Analyse the dataset imbalance problem

To study the effects of class-wise dataset imbalance on the predictive performance of the neural network model, we conducted a set of experiments by training and evaluating the classifier with different numbers and choices of target classes. Table 1 below describes the class-wise composition of training data for the conducted experiments.

Experiment ID	**Happy**	**Sad**	**Anger**	**Neutral**
1	318/318	316/316	-	-
2	318/318	316/316	-	316/787
3	318/318	316/316	107/107	-

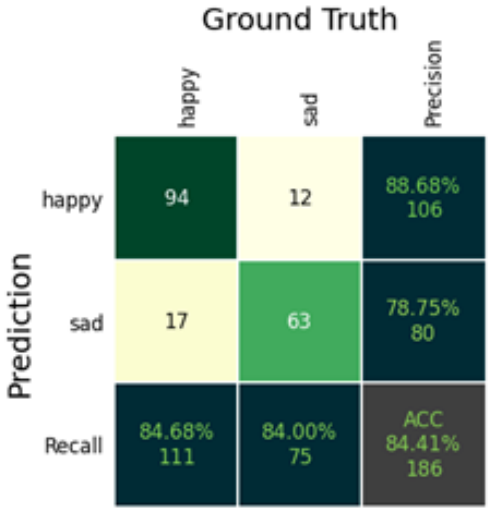
Table 1: Class wise composition (count choosen/total count) for experiments

The first experiment is conducted with the almost balanced pair of “happy”

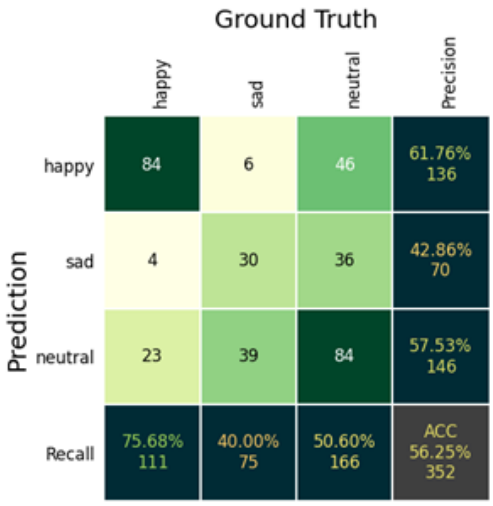
and "sad" classes, with the intent to study the effect of these classes on the predictive performance of the classifier. In the next experiment, the neutral class is additionally included after under-sampling it to match the sizes of the other two. Whereas, for the third experiment, the under-represented anger class is added along with the "happy-sad" pair to understand the bias induced by the imbalance.

The set of experiments was conducted with both "baseline" and "self-attention" variants of FAN. However, the baseline variant performed better in all conducted experiments. Models of both variants were pre-trained with Microsoft FER and AFEW datasets using the transfer learning technique. While training for the enlisted experiments, the final fully connected layer was modified in accordance with the number of target classes involved.

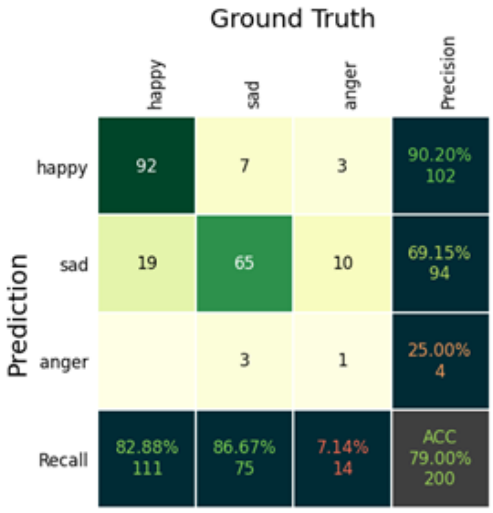
Figure 16 below contains confusion matrices that depict the performance of the baseline variant of FAN on test sets corresponding to the set of conducted experiments. The classifier exhibits a decent performance on the balanced pair of happy and sad classes with an overall accuracy of 84.41%, thereby proving its learning capacity. The high class-wise precision value associated with the happy class is an indicator of the neural net model discrimination capability. However, with the addition of the under-sampled neutral class to the scope, the overall accuracy of the classifier drops significantly to 56.25%. This strongly indicates the detrimental effect that the inclusion of a neutral class has on the model's performance. The introduction of the neutral class possibly affects the classifier's sensitivity to identify subtle emotions which make most of the frames in a video.



(a) Experiment 1



(b) Experiment 2



(c) Experiment 3

Figure 8: Confusion matrices of experiments

Unlike the inclusion of neutral class, the addition of the under-represented anger class does not drastically reduce the classifier's accuracy. However, the model performs poorly on test data from the anger class as it can correctly classify only one in a total of fourteen test videos. This is certainly due to insufficiency in the anger class's training data. Besides, most of the anger test videos are classified as sad, which is a relatively closer emotion to anger than happiness. Both the effects

discussed from the inclusion of neutral and anger classes in the second and third experiments respectively can be observed in a combined fashion from the results of the FAN model for four classes discussed in the previous subsection. It can be found that the over-represented neutral class hampers the classifier from correctly recognizing even a single test video from the anger class.

4.1.4 Qualitative Analysis

After facing the poor performance of the FAN model on target classes due to dataset imbalance, we considered undersampling the neutral class and up-sample the anger class by performing data augmentation. However, due to the inherent complexity of the augmentation task for videos, we decided to perform a qualitative analysis prior to the procedure. We qualitatively labeled every video in the anger class based on their relevance to the emotion category "only from the perspective of the visual modality". i.e. by ignoring the corresponding text and audio data for context. The videos were categorized into four relevance levels (high, medium, low, and nil) and observation comments were provided for each of them. The following Figure 9 shows a portion extracted from the excel file with labels.

Videos	Observation	Label Relevance	Prediction
007_sammer__matthias_mentalitaetsunterschiede_00127100_00146780.mp4	A subtle expression of being furious, mild frowns present across the video	High	neutral
007_sammer__matthias_mentalitaetsunterschiede_00184160_00209530.mp4	Lack of frowns but a subtle furious expression is seen. Face zoomed in.	High	neutral
009_thiemann__ellen_gescheiterter_fluchtversuch_00093861_00110920.mp4	Assertiveness expressed, fewer frowns, intense speech	Medium	anger
009_thiemann__ellen_gescheiterter_fluchtversuch_00110920_00124450.mp4	Assertiveness expressed, fewer frowns, intense speech	Medium	anger
009_thiemann__ellen_gescheiterter_fluchtversuch_00303705_00331100.mp4	Assertiveness expressed, fewer frowns, intense speech	Medium	anger
010_thiemann__ellen__einsicht_in_die_stasi-akten_00021730_00028540.mp4	Mostly neutral no frowns, poorly labeled	Low	anger

Figure 9: Labelling relavance level for angry class samples

We performed further analyses on this labeled data to better understand the nature of samples under the anger class. Analyses were performed on both the train and test splits of the anger class. However, due to the small number of samples in

the test split, we drew inferences only based on analyses performed on the train split. Figure 10 shows the composition of the train split of anger videos by label relevance and Figure 11 shows the class-wise composition of predictions for the class made by the baseline FAN model.

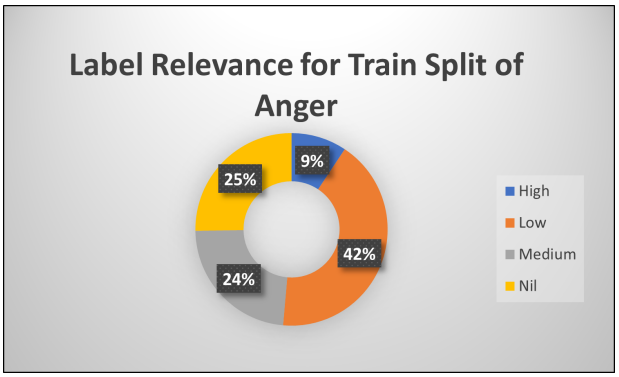


Figure 10: Label relevance for train split of anger class

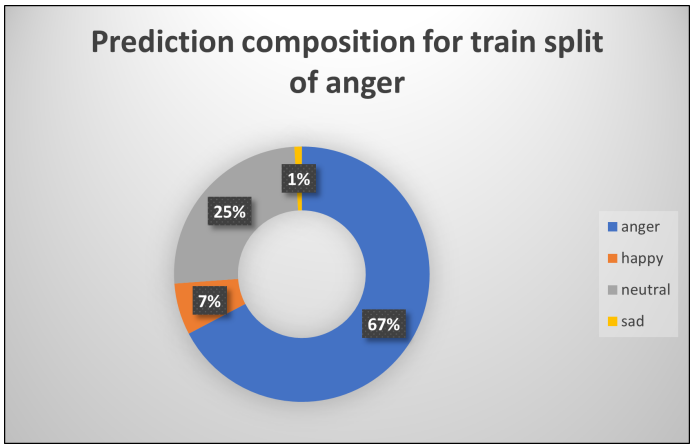


Figure 11: Class wise composition of anger predictions

Some interesting insights were obtained. It turns out that a significant portion (42%) of the videos in the anger class is less (Low) relevant to the emotion category, and about one-fourth do not fit the emotion category (classified as 'Nil'). This is due to factors such as the dominance of neutral facial expressions, the presence of a mix of expressions, and ambiguity which could be resolved only when information from other modalities are available. When it comes to the learning capacity of the

FAN baseline model, it could only fit 67% (train accuracy) of the class, leaving out one-fourth to be misunderstood as neutral. This is again due to over and under representations of neutral and anger classes respectively. At this stage, we hypothesized the presence of a significant portion of data under 'low' and 'nil' relevance categories to be the major cause for the classifier's poor performance on the already under-represented class. In order to verify this hypothesis, we created the following chart which shows the percentage of predictions by label relevance categories.

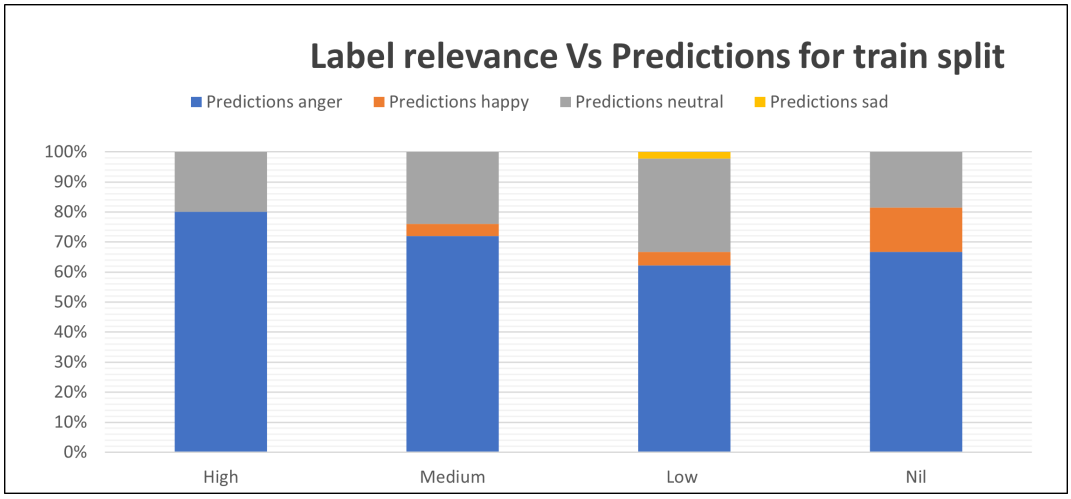


Figure 12: Predictions by label relavance categories

As hypothesized, there exists a correlation between relevance levels of samples and the FAN classifier's performance on them. About 80% of high relevance videos was fit correctly by the classifier model, while the percentage drops to 70 and 60 for medium and low, nil categories respectively. After gaining confidence from this analysis, we plotted the confusion matrix for the FAN baseline model on the entire HDG4 split as shown below in Figure 13.

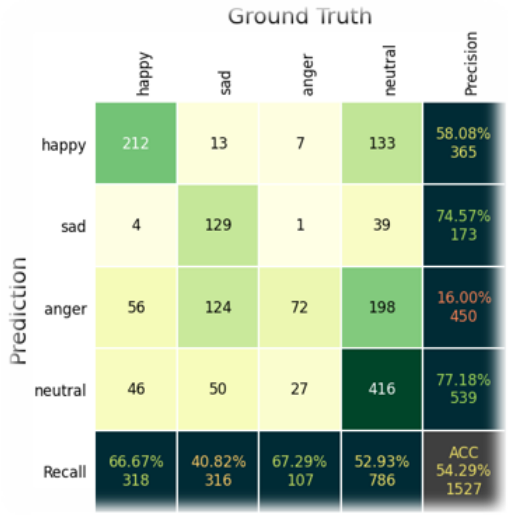


Figure 13: Confusion matrix for FAN baseline model on full HDG4 train split

It turns out that the problem extends beyond the anger class to others. It's more evident from the training recall of the sad class (40.82%) and the over-represented neutral class (52.93%) that ambiguity in labeling needs to be addressed to improve the performance of the classifier.

4.1.5 Future work

The future work should focus on performing following experiments to improve the performance of the classifier:

- Data-augmentation
- Use a weighted loss function to handle dataset imbalance
- Reformulate the task from single to multi-label classification to address the labeling ambiguity problem.
- Hyper-parameter tuning and architectural changes such as the introduction of multi-head attention.

4.2 Krishna's experiments

This section discusses the continuation of work done by Aswin Kumar and focuses on solving the problem by considering the suggested future work in Section 4.1.5.

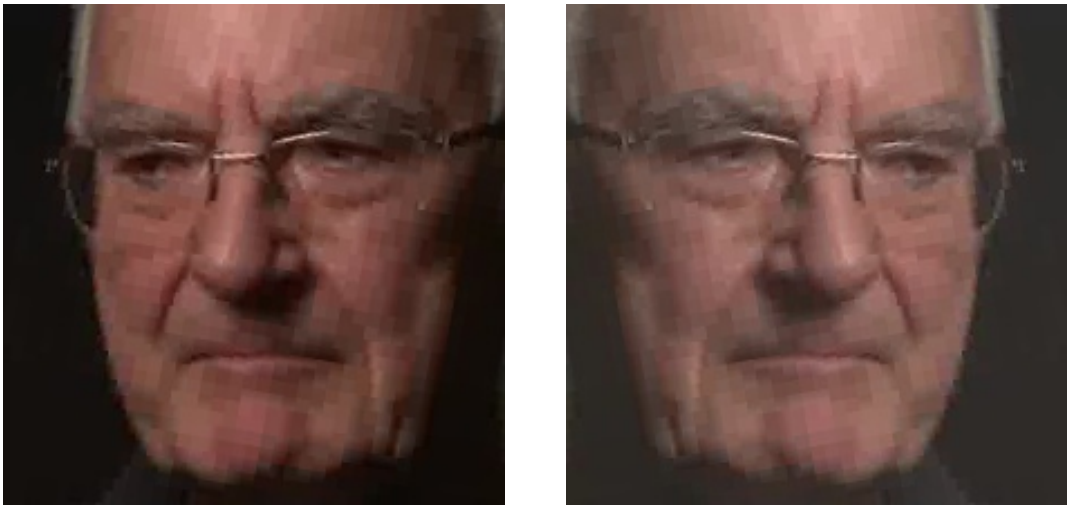
4.2.1 Handling Imbalanced problem :

From the previous experimental results, we can infer that the dataset is imbalanced and labeling ambiguity affects the performance of the model. Generally, the imbalanced dataset problem is handled using 3 techniques [5]:

- Data level techniques - Undersampling, oversampling
- Algorithm level techniques - weighted loss functions
- Hybrid techniques - Combination of data and algorithm level techniques

Therefore, the experiments performed in this section makes use of these techniques to investigate their effect on the performance of the classifier.

Data Augmentation :
Data augmentation is a task to increase the size of the dataset by adding modified data from the original dataset or creating synthetic data from the original dataset. Data augmentation falls under the Data level technique to solve the problem of imbalanced datasets [6]. Data augmentation is generally performed using traditional and Deep learning(DL) based methods [6]. Generative Adversarial Networks(GANs) are believed to be state-of-the-art DL techniques for augmenting the dataset. However, GANs technique is unreliable and losses temporal relation between frames in the video [6]. In order to preserve the temporal relation, a uniform transformation should be performed across all the frames of a video [6]. Therefore, traditional augmentation methods such as Horizontal flip, brightness, and contrast transformations are performed on all video frames. This augmentation is performed only on the Angry class of HDG4, which is the minority class in the dataset. Figures show the original and augmented image of the angry class. The original dataset has an imbalance ratio of 4.08 on the train set and 8.04 on the test set. After augmentation, the imbalance ratio is 2.04 on the trainset.



(a) Original image (b) Augmented

Figure 14: Augmentation of Anger class

Weighted Loss fuction :

The weighted loss function is another solution for handling imbalanced dataset problems. Here, the minority samples are given more weight for contributing to the model loss. Therefore the model tries to learn these hard misclassified samples. This method falls under the algorithm level category for solving the imbalanced problem. Focal loss is the most prominent and state-of-the-art cost function for handling imbalanced datasets [7]. Table 2 contains the weights associated with each class of HDG4 in the loss function. From the Equation below, the focal loss has a modulating parameter gamma, which is between 0 to 5. For our experiment's gamma value, 2 is used as suggested by the paper [7].

$$FL(P_t) = -(1 - p_t)^{\gamma} log(p_t) \tag{1}$$

Class	Total samples	weight
Anger	44202	2.459
Happy	127110	0.855
Sad	83699	1.298
Neutral	179680	0.605

Table 2: Class Weights for focal loss function

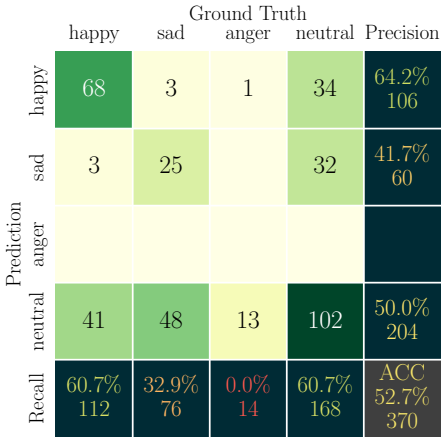
Experiments :

The following experiments are performed to analyze the performance of the FAN baseline model on HDG4 dataset.

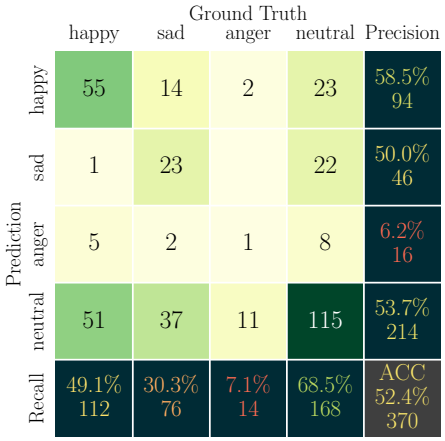
ID	Experiment name	Happy	Sad	Neutral	Anger	Total videos
1	Baseline	318	316	786	107	1527
2	Baseline_anger_aug	318	316	786	214	1634
3	Baseline_undersampled	214	214	214	214	856
4	Baseline_focal_loss	318	316	786	107	1527
5	Baseline_focal_loss_aug	318	316	786	214	1634

Table 3: Class wise composition of HDG4 train split for experiments

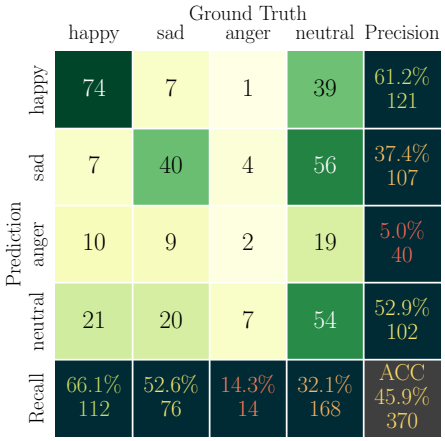
The first experiment uses the full train split for training the FAN. The second experiment training data contains an additional number of anger videos due to the augmentation of the anger class. The third experiment training data consists of the same number of videos across all classes by undersampling the neutral, sad, and anger classes to the size of the augmented anger class. In the fourth experiment, the baseline model is trained with focal loss by using the corresponding weights given in Table 2. The fifth experiment uses the hybrid technique to solve the imbalance problem by utilizing both augmentation and weighted loss function.



(a) Exp 1: Baseline

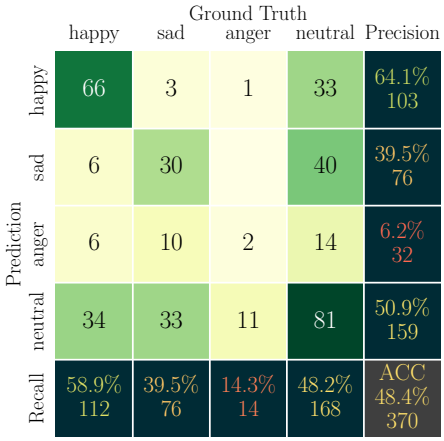


(b) Exp 2: Baseline_anger_augmented

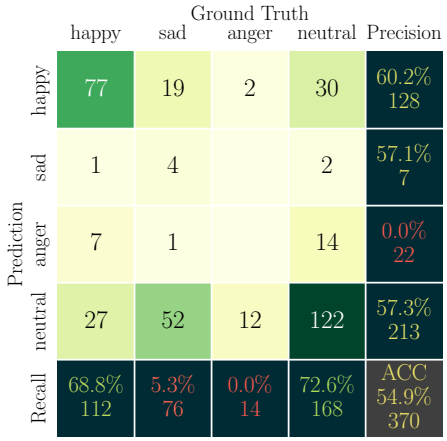


(c) Exp 3: Baseline_undersampled

Figure 15: Confusion matrices of experiments performed using FAN baseline model



(a) Exp 4: Baseline_focal_loss

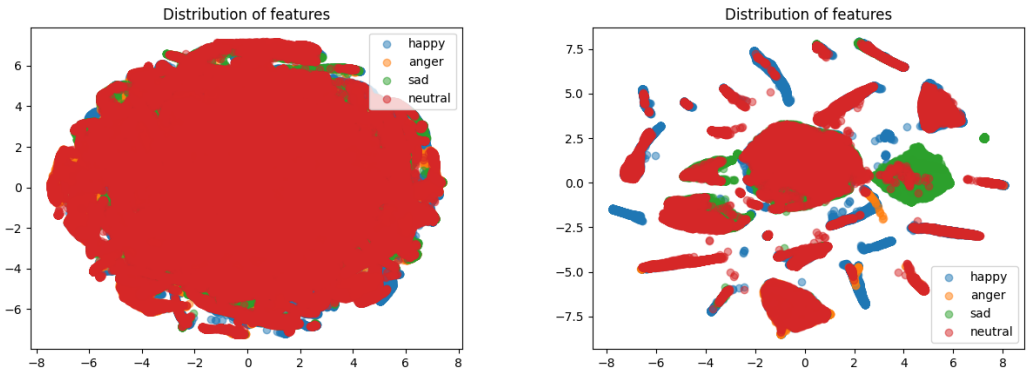


(b) Exp 5: Baseline_focal_loss_augmented

Figure 16: Confusion matrices of baseline model with focal loss

- From the experimental results, the augmentation of the anger class improved the model's ability in trying to predict the anger class, which can be seen in Figure 15b. But the overall performance of the models hasn't changed much compared to the model trained using the original HDG4 dataset.
- The undersampling of the dataset to the size of the augmented anger class didn't have much impact on the overall improvement of performance, even though the dataset is not imbalanced.
- The baseline model with focal loss resulted in less accuracy, but the model is trying to predict the minority samples. This explains that the weighted loss function improved the model's ability to predict minority class labels when compared to the baseline model.
- In the third experiment, the model is trained using more samples of anger class and focal loss function (Hybrid technique), which also didn't bring much of an improvement in the performance of the model.
- From the previous student work, the self-attention model didn't improve the performance. Therefore, similar experiments on self-attention and relation attention models are not performed.

From the observed experimental results, it is evident that imbalance in the dataset is not the only reason for the performance but ambiguity in the dataset. The feature space of the model helps in understanding how the features are distributed. Therefore, using the dimensionality reduction technique TSNE, the feature space is visualized on a 2D plane. The below Figure 17 shows the distribution of features.

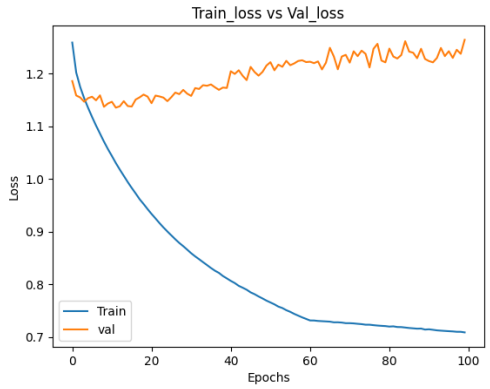


(a) Train split feature plot

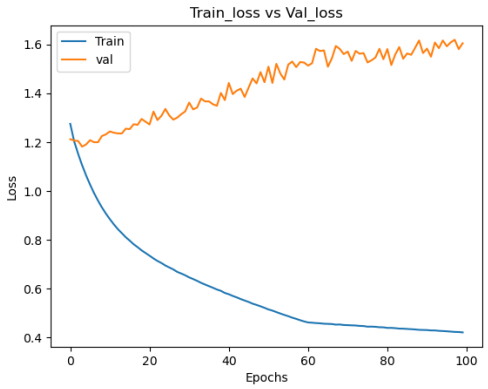
(b) Test split feature plot

Figure 17: 2D Feature space plot of HDG4 dataset

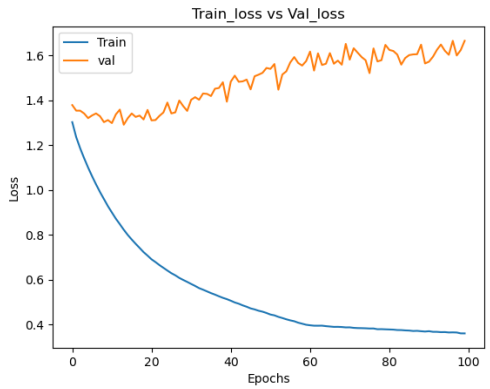
- From Figure 17a, the neutral class features are spread out across the feature space and overlapping over all the remaining emotion classes.
- However, in the test split from Figure 17b, the happy and sad features are not much overlapped by the neutral class. This explains the precision and recall of both happy and sad classes by the model when compared to the anger class.
- From the above experimental results and these feature space plots, we can infer that all the emotion class features are highly correlated and thus affecting the performance of the model even after performing augmentation and weighted loss functions.
- Additionally, from the training vs validation plots in Figure 18, we can infer that all the models are overfitting. Which is also obvious from the experimental results.



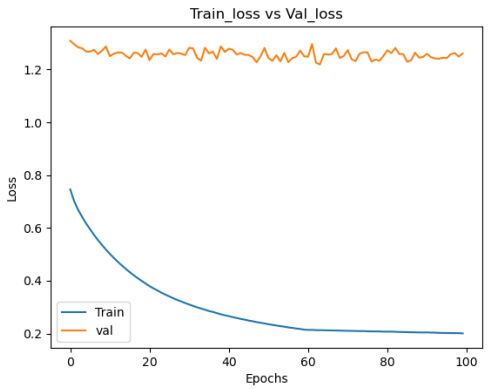
(a) Exp 1: Baseline



(b) Exp 2: Baseline_anger_augmented



(c) Exp 3: Baseline_undersampled



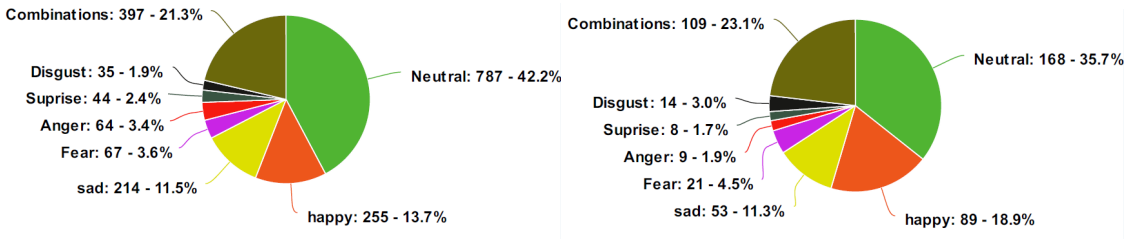
(d) Exp 4: Baseline_focal_loss

Figure 18: Train vs Validation loss plots of the experiments

4.2.2 Multi-label classification:

From the previous experimental results, we can infer that the dataset has a labeling ambiguity problem. Therefore, this might be solved by converting the Multi-class classification task to a multi-label classification task, which is also suggested in the future work of Ashwin Kumar. In multi-label classification, the model predicts more than one emotion class for a video instead of a single class in the case of multi-class classification. The annotation for multi-label dataset is done by Viswanath Anargh. HDG multi-label dataset has 7 classes namely, Happy, Sad, Anger, Surprise, Disgust, Fear, and Neutral. This particular dataset is referred to as "HDG_ML" in this report. For example, now the ground truth of the label looks

like 1000100, which means the video contains both Happy and Disgust emotions (label bits corresponds to the order of the 7 classes mentioned). The dataset and its class composition can be observed from Figure 19.



(a) Train split (b) Test split

Figure 19: Class distribution of HDG multi-label dataset

The metrics used for Multi-label classification are Accuracy (Exact match ratio), Hamming loss, Precision, Recall, and F1 score [4]. Accuracy in this case considers all the instance's predictions to match the ground truth instances. whereas, the hamming loss is calculated by considering partial correct predictions. Hamming loss value ranges between 0 and 1. Nearer to 0, the better the model performance in classifying the dataset. Table 4 illustrates the experiments performed to train the FAN baseline model on the multi-label dataset.

Exp ID	**Neutral**	**Happy**	**Sad**	**Fear**	**Anger**	**Surprise**	**Disgust**	**Others**	**Total videos**
1	787	255	214	67	64	44	35	397	1863
2	-	255	214	67	64	44	35	397	1076
3	787	-	214	67	64	44	35	397	1608
4	787	255	-	67	64	44	35	397	1649
5	255	255	214	67	64	44	35	397	1331

Table 4: Experiments for Multi-label classification and their train split data distribution

The first experiment uses the full train split of the HDG_ML dataset to train FAN baseline model. The remaining experiments in the table are performed to analyze the effect of class wise imbalance on the predictive performance of the

model. Therefore, the second experiment doesn't include neutral class samples for the training of the model. Similarly, the third and fourth don't include happy and sad samples respectively for training the model. In the fifth experiment, neutral and happy class samples are under sampled to the size of 255 videos which is very close to the samples of the sad class.

Experimental Results:

ID	Experiment name	Precision	Recall	F1 score	Hamm loss	Accuracy
1	Baseline	0.418	0.437	0.427	0.216	0.303
2	Baseline_no_neutral	0.524	0.485	0.504	0.248	0.273
3	Baseline_no_happy	0.392	0.423	0.407	0.234	0.313
4	Baseline_no_sad	0.46	0.476	0.468	0.208	0.33
5	Baseline_N_H_S	0.329	0.315	0.321	0.254	0.175
6	Self_Attention	0.22	0.201	0.212	0.203	0.176
7	Baseline_HDG4	0.471	0.503	0.486	-	0.502

Table 5: Experimental results

- Table 5 illustrates the results form the various experiments performed on FAN to classify the dataset.
- The results from baseline model(experiment1) trained on HDG_ML showed that the performance of model haven't improved much when compared to the baseline model(exp7) performance on HDG4 dataset, which is multi-class dataset. The F1 score of baseline model on HDG_ML dataset recored 0.427. whereas, baseline model on HDG4 dataset recorded 0.486.
- Therefore, other experiments are performed to analyse the effect of majority classes on the predictive performance of the model. The majority classes are happy, neutral and sad.
- Baseline model trained without neutral class in the dataset improved the models f1sore from 0.427 to 0.504. This is also the model with highest f1score among all other experiments. Neutral class holds 787 videos out of 1863

videos in the dataset. This strongly indicates the effect of neutral class on degrading the models performance.

- Even after reformulating the task from single to multi-label classification to address the ambiguity by re-annotating the dataset, the effect of imbalance still exists. By comparing the data distribution between the HDG dataset in Figure 4 and HDG_ML dataset in Figure 19, it is evident that the percentage of samples that belong to neutral class are same in both train and test split of two datasets. Meanwhile, other classes samples fall under the combinations, where video contains more than one emotion class.

- From the experiment 3, we can infer that happy class has not much effect on the performance of the model. Whereas, removing sad class from dataset(exp 4) improved the model performance by 4%.

- Undersampling the neutral class to 255 samples, degraded the model's F1 score to 0.321. Experiment with no neutral class have same number of happy and sad samples, which means adding 255/787 neutral class samples degraded the performance of the model. Similar phenomenon can be observed from the initial experiments in Section 4.1.3, confusion matrices of experiments in Figures 8a, 8b. where, the addition of under sampled neutral class degraded the performance of model. Therefore, the neutral class effects the overall discrimination capacity of the model to predict subtle expressions.

- self-attention variant of FAN is trained on full HDG_ML dataset and resulted in F1score of 0.212. Due to the correlation between the features of emotions, learning of attention weights are misguided and failed to classify on test set.

5 Conclusion

This research work focuses on developing a video-based emotion recognition, which intends to detect the emotions present in the whole video of oral history interviews from Haus der Gesichte (HDG). The project uses dataset which consists of few seconds cut of whole videos from HDG, where videos are classified based on the predominant emotions expressed by the subjects present in the video. we conducted a literature review for video based Facial expression recognition and choose Frame attention networks in this project, which is based on attention mechanism. HDG dataset originally consists of 7 emotion classes but initial experiments are performed considering only 4 main emotions Happy, Sad, Anger and Neutral. From the experimental results we can infer that the predictive performance of FAN on anger class is effected by its under representation and over representation of neutral class in the dataset. Further qualitative analysis proved that the ambiguity in the dataset labelling needs to be addressed to resolve the classification performance. Additionally, experiments are performed to solve the imbalance in dataset by undersampling, over sampling, weighted loss functions and augmentation. However, the performance of model has not improved. Therefore, to understand deep about the emotion classes, feature representations of classes are plotted using dimensionality reduction technique Tsne. The feature plots showed that the emotion classes share feature similarity with neutral class. Thus effecting the FAN predictive performance. Single label task is reformulated to multi-label classification task to study the ambiguity. However, the results are not as expected due to the imbalance correlation in dataset. Additionally, from all the results it is pretty clear that the emotion classes are not having enough features from just single modality to predict the true emotion. Thus the multiple modalities are required to detect the true emotion. Therefore, the future work should focus on processing the multi-modalities to solve the ambiguity problem and improve the predictive performance of model.

References

[1] Jingjun Liang, Li Ruichen, and Qin Jin. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. pages 2852–2861, 10 2020.

[2] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. pages 2401–2410, 01 2021.

[3] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. pages 3866–3870, 09 2019.

[4] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, PP, 04 2018.

[5] Justin Johnson and Taghi Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27, 03 2019.

[6] Nino Cauli and Diego Reforgiato Recupero. Survey on videos data augmentation for deep learning models. *Future Internet*, 14:93, 03 2022.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. pages 2999–3007, 10 2017.