Introduction

The idea of the project is to help detect audience engagement by identifying the audience's mood by capturing their faces. This could enable the speaker to get real-time feedback on people's reactions. This can also be extended to let the speaker know the speaker's impact on the audience, such as how many people feel happy after narrating a feel-good story. Especially in a virtual setting, a speaker might lack such feedback, which could help the speaker dynamically adjust to increase the audience's engagement and give a better overall experience to the listeners. In order to achieve this, we will be relying heavily on the paper, "Facial Expression Recognition in the Wild via Deep Attentive Center Loss," which is a recent and highly accurate way for facial expression recognition (FER) and has been proven to work on varied emotion datasets such as RAF-DB and AffectNet. This paper brings in the Deep Attentive Center Loss (DACL) method, which helps tackle the FER problem better than old methods. The DACL method has been proven to work well on varied emotion datasets like RAF-DB and AffectNet, and should be potentially good in identifying emotions from images taken in the wild (Non ideal or laboratory scenarios).

Background and related work

The different approaches and techniques that have been used to improve facial expression recognition (FER) can be broadly viewed from two different angles, methods using Deep metric learning (DML) and methods that tackle the challenges posed by wild FER datasets.

Let's first discuss FER with DML, which essentially involves using DML techniques to improve the performance of FER models. These methods aim to better handle the large intra-class variation and inter-class similarity that can make FER challenging. Some specific techniques include the use of contrastive loss, (N+M)-tuplet cluster loss, and Locality-Preserving loss.

The second one specifically tackles the challenges posed by wild FER datasets, which refer to datasets with diverse subjects in unconstrained environments. Some specific challenges include face occlusion, pose, and label scarcity. Techniques here include the use of attention mechanisms to filter out irrelevant features or patches, the use of region attention networks to handle occlusion and pose, and the use of semi-supervised learning and inductive transfer learning to address label scarcity.

Methods

So this project aims to test how well facial recognition models can perform in real world in terms of accuracy. Although the efficiency of the model is crucial to having real time feedback of the audience, which requires the model to process through a high number of frames to make predictions, we mainly focused on the accuracy due to the nature of computation getting vastly better over the years.

In our case, to actually test the model which we build around the ResNet 18 architecture and the techniques such as Deep Attentive Center loss, which is mentioned in the paper, "Facial Expression Recognition in the Wild via Deep Attentive Center Loss", we wanted to use some personally made webcam videos and see how well the model performs.

About the model:

We follow the method from the paper, “Facial Expression Recognition in the Wild via Deep Attentive Center Loss” very closely for model architecture and training.

The first step includes training a mini-batch of m samples for a K-class classification problem. A CNN generates a spatial feature map, and a pooling layer extracts a d-dimensional deep feature xi. A fully-connected layer with class weights and bias parameters maps the deep feature to a raw score vector, and a softmax function calculates the probability distribution p(y=j|xi) over all classes. The cross-entropy loss function measures the discrepancy between prediction and the true label to formulate the softmax loss function as below,

$$\mathcal{L}_S = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{K} y_i \log p(y=j|x_i)$$
$$= -\frac{1}{m}\sum_{i=1}^{m} \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{K} e^{w_j^T x_i + b_j}}$$

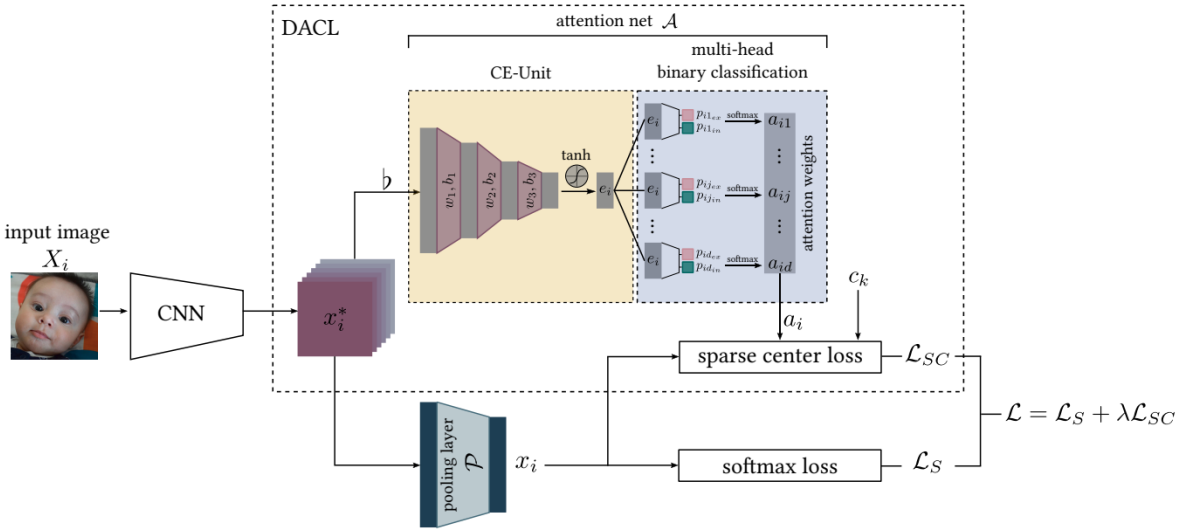
Sparse Center Loss

The center loss is a Deep Metric Learning method that measures similarity between deep features and their class centers. It minimizes the Within Cluster Sum of Squares (WCSS) to partition the embedding space into K clusters for a K-class classification problem. It aims to minimize the difference between the deep feature vector and its corresponding class center. To minimize the Within Cluster Sum of Squares (WCSS), the penalty for the Euclidean distance between the deep features and their respective class centers in the embedding space is applied equally.

$$\mathcal{L}_C = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{d} \|x_{ij} - c_{y_i j}\|_2^2$$

Sparse center loss makes the process even more efficient by selecting only the relevant elements in a deep feature vector for discrimination. This is done by weighting the Euclidean distance calculated at each dimension and developing a sparse center loss method. The goal is to filter out irrelevant features in the discrimination process.

$$\mathcal{L}_{SC} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij} \odot \|x_{ij} - c_{y_i j}\|_2^2$$



Attention Network

Attention network is attached to the CNN to estimate the weights for the sparse center loss based on the input. The weights are determined by the neural network, which adaptively computes an attention weight vector. The proposed attention network consists of two main components: the Context Encoder Unit (CE-Unit) and the Attention Estimator Unit (AE-Unit). These two components work together with the sparse center loss to form the proposed DACL method. Figure 2 shows the structure of the attention network.

The Context Encoder Unit (CE-Unit) takes the CNN spatial feature map as input and generates a latent representation, and the multi-head binary classification module that estimates the attention weights. The CE-Unit includes three trainable fully-connected linear layers to extract only the relevant information from the context. The context for the attention network is at the convolutional feature-level to preserve the spatial information. The module calculates two raw scores, one for inclusion and the other for exclusion, for each dimension in the deep feature vector.

$$p_{ij_{in}} = A_{j_{in}}^T e_i + b_{j_{in}}$$
$$p_{ij_{ex}} = A_{j_{ex}}^T e_i + b_{j_{ex}}$$

The softmax function is applied to normalize the scores, and the corresponding attention weight is calculated.

$$a_{ij} = \frac{\exp(p_{ij_{in}})}{\exp(p_{ij_{in}}) + \exp(p_{ij_{ex}})}$$

Training

The proposed DACL method is trained in an end-to-end manner by jointly supervising the sparse center loss with the softmax loss to create the final loss as below,

$$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_{SC}$$

λ is used to control the sparse center loss contribution. The parameters can be optimized using the Standard Gradient Descent.

The gradient of sparse center loss with respect to Deep Features

$$\frac{\partial \mathcal{L}_{SC}}{\partial x_i} = \frac{1}{m} a_i \odot (x_i - c_{y_i})$$

The gradient of sparse center loss with respect to attention weights

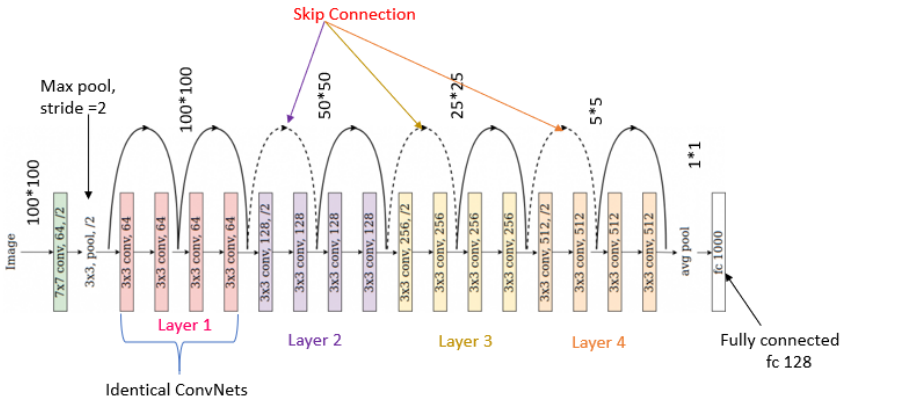
$$\frac{\partial \mathcal{L}_{SC}}{\partial a_i} = \frac{1}{2m} \|x_i - c_{y_i}\|_2^2$$

Training Data Set and Techniques:

We'll be using the RAF-DB dataset to train our model. The decision was made based on the availability of the data sets and the number of training images available that could be effective for the model.

The Real-world Affective Faces Database (RAF-DB) is a facial expression database consisting of around 30,000 images with diverse characteristics such as age, gender, ethnicity, head poses, lighting conditions, and occlusions. Each image in the database has been labeled by approximately 40 annotators, and the annotations include a 7-dimensional expression distribution vector, landmark locations, bounding box, race, age range, and gender attributes. The database contains two subsets: a single-label subset with 7 classes of basic emotions, and a two-tab subset with 12 classes of compound emotions. Additionally, the database includes baseline classifier outputs for both basic and compound emotions.

We used PyTorch framework to aid us in the project. We will be comparing the method of training with pretrained weights on the ResNet18, trained by ImageNet DB and without any pretrained weights to see if that makes any difference and would use the one that has a better validation accuracy.



ResNet-18 Architecture

Fruit 360 Input Image size= 100*100 px

Extracting important frames from Videos:

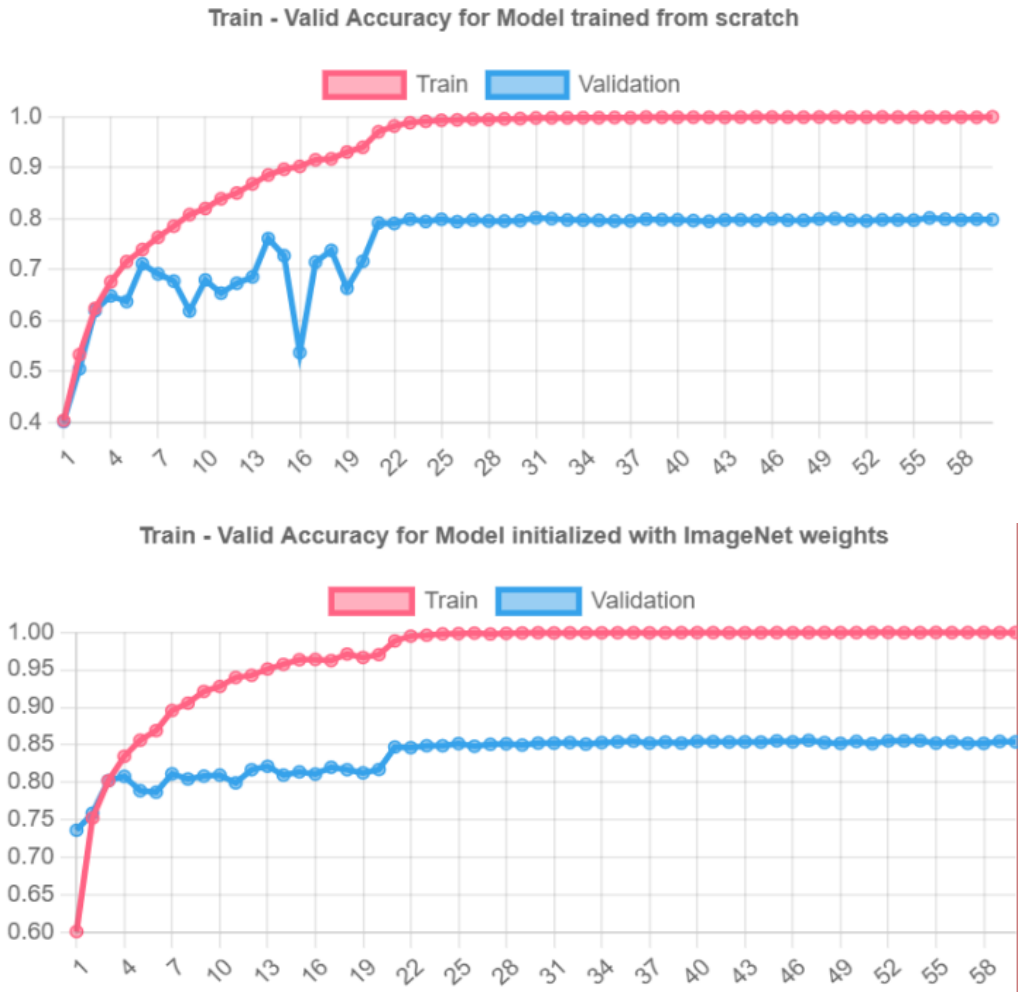
Our project has a requirement to process videos and then predict the emotion from them. It could get computationally very intensive to actually predict through all the frames in a video as a typical video has 30 frames per second. It is important that only the frames, which are

relevant are filtered and sent through for prediction. In order to do that, we select only the frames which have a probability of change in the scene as that could denote a change in the emotion of a face in the video. To implement this, we used FFmpeg, which is a free and open-source software project that offers many tools for video and audio processing. We experimented with the probability threshold value to detect scene change with face web cam videos and set it as 0.0030.

Experiment:

Our model was trained and evaluated for a total of 60 epochs, using both a pretrained model on ImageNet and starting from scratch. Each run of the training process took approximately 12 hours. To organize the data for training and validation, we categorized the images in the Real-world Affective Faces Database (RAF-DB) into 7 distinct classes based on their annotations

We compared the results from training with pre-trained weights on the ResNet18, trained by ImageNet DB and without any pre-trained weights. It was observed that the model with pre-trained weights performed better with the validation test. We could see a significant (around 5%) jump in the validation accuracy as shown below,



Results

Best validation epoch result for Resnet18 trained from scratch

```
[>] Best Valid --------------------
[+] acc=0.8012
[+] rec=0.6919
[+] f1=0.7069
[+] aucpr=0.7210
[+] aucroc=0.9331

[*] Fini! ------------------------
[!] total time = 11:42:27.499749s
```

Best validation epoch result for Resnet18 trained from pretrained ImageNet weights.

```
[>] Best Valid --------------------
[+] acc=0.8559
[+] rec=0.7622
[+] f1=0.7745
[+] aucpr=0.8184
[+] aucroc=0.9624

[*] Fini! ------------------------
[!] total time = 12:03:43.701703s
```

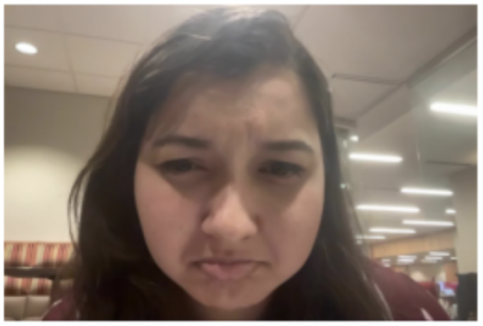
Some of the validation data results are:

Since the aim of our project was to detect and quantify the facial expressions of people in a real life audience engagement setting like a video conference, we captured videos of different people making facial expressions through our webcam to mimic such an event, and ran them through an algorithm to detect keyframes from the videos, predict the emotions in each of the key frames, and finally quantify them as an aggregation of emotions through the entire video as a percentage measure. This will give us a snapshot of how engaged the audience were throughout the event

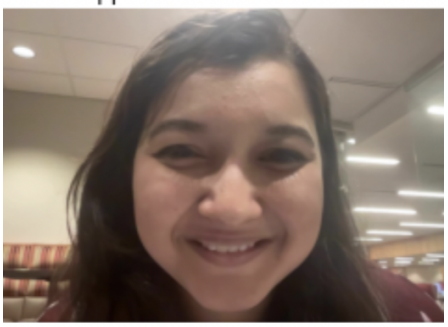
Some of the results when we tested the model's performance against real life settings through our webcams:

Results that were close to ground truth:

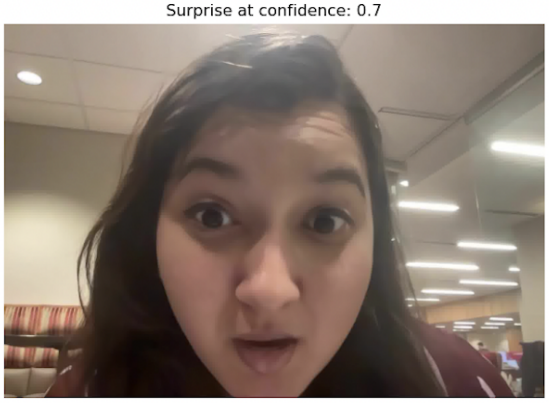
Sadness at confidence: 0.58



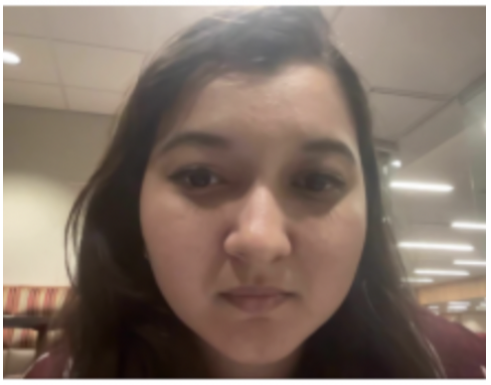
happiness at confidence: 0.33



Surprise at confidence: 0.7

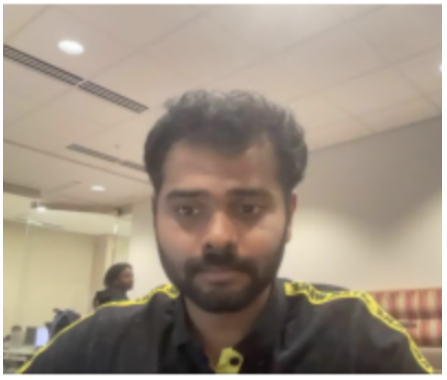


Neutral at confidence: 0.67



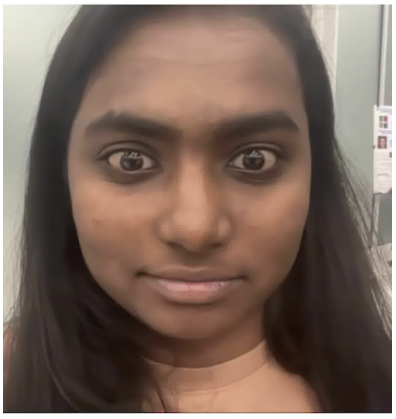
Surprise at confidence: 0.65

Neutral at confidence: 0.61



Sadness at confidence: 0.51

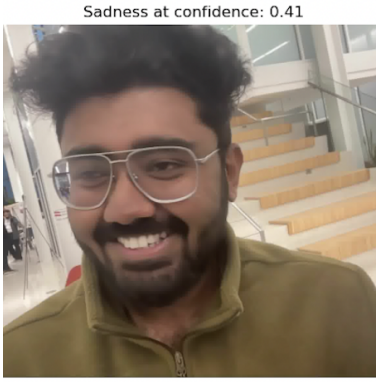
Surprise at confidence: 0.53



happiness at confidence: 0.4

Surprise at confidence: 0.74

Incorrect predictions:



Discussion

The model performed well with the data from the RAF-DB dataset, when we sampled images from the test data set and ran the model multiple times. It reached around 85% with the pretrained models in 60 epochs, which is in line with the expectations set by the paper.

Limitations:

Since we were testing the model through homemade webcam style videos, it was difficult to quantify the percentage of accuracy due to the random way of testing and no fixed frame selection from the videos. But with the heuristic analysis, we did find that the model struggled with homemade videos. It had a skew towards detecting neutral as the emotion, and the model could almost never predict “disgust” as an emotion.

Preprocessing could have been the issue, since the videos we tested were taken through webcams, to simulate a real life conference scenario, and there was no fixed transformation that could exactly align with the training images. The problem was mostly with the resizing and normalizing values we used for the keyframes from our videos. We tried tweaking some parameters to increase the accuracy, which helped a little, but there is still room for improvement.

Conclusion

The current model we have developed has shown promising results with an accuracy of approximately 85% in detecting emotions from facial images, which is impressive

considering that we only used the RAF-DB dataset. However, we recognize that further improvement is necessary, and training the model with other FER datasets such as AffectNet may lead to higher accuracy. Nonetheless, it should be noted that this result was only achieved with the validation dataset. When tested on our homemade webcam-style videos, the model encountered difficulties in accurately detecting emotions and displayed a bias towards the neutral emotion. This discrepancy may be due to the training images showing emotions in a more exaggerated way.

Despite the challenges in accuracy with respect to real-world videos, the project can seed the solution of enabling speakers, especially in a virtual setting to get some feedback on the emotional response of the listeners, which is crucial for the speaker to adapt dynamically. Also the benefits of the project extend to areas like, Human-computer interaction, where enabling devices to respond to human emotions or expressions is very intuitive, and also in areas like marketing and advertising, where analyzing customer reactions to marketing materials can improve marketing strategies.

To achieve better accuracy, future work entails integrating more diverse and accurately annotated data from natural settings, which poses a challenge due to the sensitive nature of the data and the manual effort required for annotation. One potential solution is to use GAN-generated images, although this approach may introduce additional complications and we can know if that works only after further testing. Additionally, the task requires optimizing the algorithm's efficiency and utilizing high-quality hardware to provide real-time emotion detection for high-quality videos, which is crucial in real-world applications.

Source Code and Materials:
https://drive.google.com/drive/folders/1xe4Bd1O3BKiUhV1aAtVoMu2MwcocPudY

References

- Farzaneh, Amir Hossein, and Xiaojun Qi. "Facial expression recognition in the wild via deep attentive center loss." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021.
- Li, Shan, Weihong Deng, and JunPing Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- Wang, Kai, et al. "Region attention networks for pose and occlusion robust facial expression recognition." *IEEE Transactions on Image Processing 29 (2020): 4057-4069.*