MQM 546Q: Modern Analytics

FACIAL EMOTION ANALYSIS IN CANDIDATE EVALUATIONS

Team 60, Section C

Tim Gong, Caterina Mora,

Fariha Tasnim Roza, Koko Wan, Priyesh Singh

1. MOTIVATION AND DATA UNDERSTANDING

Understanding emotions from facial expressions is a useful task with widespread applications in business. The ability to detect emotions and understand these can improve interactions and clarity adding a deeper level of connection in hiring processes. Automating this task using computer algorithms not only enhances efficiency but also mitigates human biases, providing objective and consistent results. This project aims to leverage deep learning techniques to classify facial expressions into seven emotions: angry, disgust, fear, happy, sad, surprise, and neutral.

The primary business application we will focus on are recruitment assessments and interviews, where emotional analysis can provide insights into candidates' engagement, stress levels, and reactions during the hiring process. For example, companies in investment banking or consulting use assessments that show scenarios in an interactive gamefield to assess decision-making and responses to dynamic situations. Adding live emotion recognition during these game assessments would help evaluating emotional resilience and understand stress and decision making more in depth. For example, if a candidate shows consistent high stress (e.g. angry, disgust, and fear) during case questions or technical problem-solving tasks, the system could flag this as an area for further evaluation. Other uses could be to detect culture fit, monitoring candidate engagement over time, and identifying comfort zones. This technology could provide more information into the decision-making process and provide unbiased information, improving efficiency.

We used 3 datasets from Kaggle, "FER-2013", "Affectnet Database", and "MMA FACIAL EXPRESSION". "FER-2013" consists of a total of 35,887 images: 28,709 in the training set and 7,178 in the test set. The images are 48x48 pixels grayscale, and the faces have been already centered and standardized. "Affectnet Database" has a total of 34,553 images: 25,124 in the training set and 9,429 in the test set. Images are 96x96 RGB. "MMA FACIAL EXPRESSION" has a total of 127,680 images: 110,324 in train and valid combined, and 17,356 in test. Images are 48x48 with a mix of grayscale and RGB. All photos from these three datasets have already been centered and standardized.

2. DATA PREPARATION

The structures of the three datasets differ slightly. “FER-2013” and “MMA FACIAL EXPRESSION” have the same seven different classes of expressions: anger, disgust, fear, happy, sad, surprise, and neutral. “Affectnet Database” has one more class, contempt, which we excluded to maintain consistency across the datasets. We found it difficult to clean and combine all the three datasets together directly with Kaggle API and load them to Google Colab. Therefore, we first cleaned the datasets locally and reuploaded them to Google Colab.

Next, we preprocessed all the images to prepare them for training. Since we are going to use the Vision Transformer (ViT) model later, which requires input images in RGB format with dimensions of 224x224, we standardized all images to match this configuration. To enhance the variability of the training dataset, we applied data augmentation techniques, including random flips and rotations. We checked classes across the three datasets to make sure they are the same before proceeding. After combining the datasets, we obtained a total of 164,157 images for training and 33,963 images for testing. The images are not distributed evenly across emotion classes. As shown in figure 1, we have considerably more images in happy and neutral classes. Additionally, we split the training data into training and validation sets using an 80:20 ratio, ensuring that the model is evaluated on unseen samples during training for better generalization.

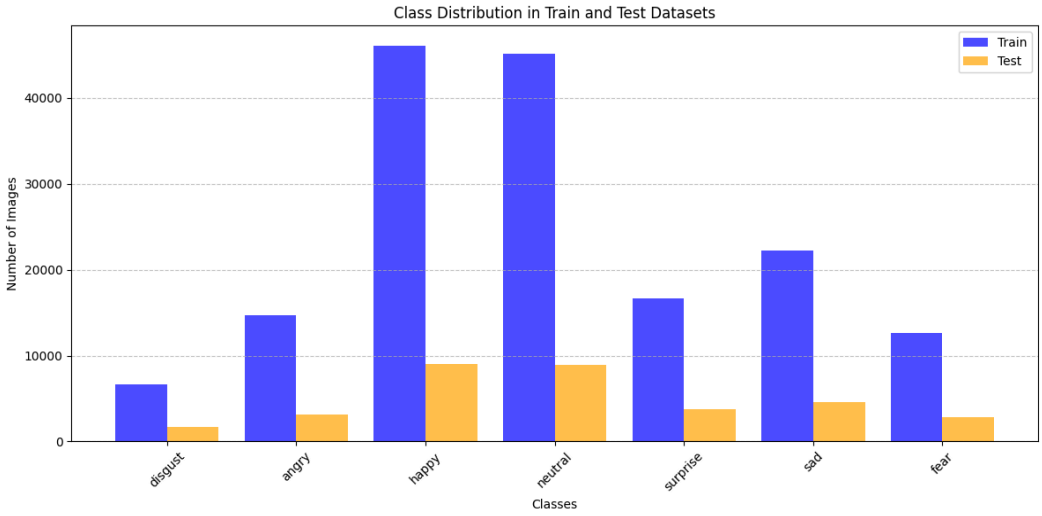

Figure 1: Class Distribution of Emotions in Train and Test Datasets

Lastly, we created data loaders to efficiently handle model training and evaluation. We set the batch size to 32, enabling the model to process a manageable group of images at a time. The following figure illustrates a sample batch of 20 images from the training set.

Figure 2: 20 Sample Images in Train Set

3. MODELING

Our modeling approach employs a Convolutional Neural Network (CNN) to classify emotions directly from facial imagery. The CNN architecture leverages convolutional layers to extract features such as edges and textures, pooling layers to condense spatial information, and fully connected layers to transform these features into emotion predictions. By training with CrossEntropyLoss and optimization algorithms like SGD or Adam, the model iteratively refines its parameters over multiple epochs. Validation loss tracking ensures that the model generalizes effectively to unseen data. Through this structured, hierarchical learning process, the CNN becomes adept at associating nuanced facial features with distinct emotional states.

Baseline Model

A baseline Convolutional Neural Network (CNN) model was developed and evaluated on the FER-2013 dataset, which comprises seven emotion categories. The model was trained for 30 epochs and tested on approximately 7,000 images, achieving an overall accuracy of 57%.

Architecture:

- **Convolution & Pooling:** Three convolutional layers each followed by 2x2 max-pooling. Feature maps:
    - Conv1: 32x32x3 → 16 filters → Pool → 16x16x16
    - Conv2: 16x16x16 → 32 filters → Pool → 8x8x32
    - Conv3: 8x8x32 → 64 filters → Pool → 6x6x64

- **Classification Head:** Flattened output → FC (500 neurons, ReLU) → Dropout (0.25) → FC (7 classes).

Results:

"Happy" (82%) and "Surprise" (70%) were well-classified, while "Angry," "Neutral," and "Sad" hovered around 50%, and "Fear" and "Disgust" remained at 23%. Misclassifications stemmed from similar facial features and class imbalance.

Challenges & Improvements:

Misclassifications stemmed from overlapping features and class imbalance, especially for "fear" and "disgust." Improving data diversity, exploring deeper architectures (e.g., ResNet), and fine-tuning hyperparameters could enhance overall performance.

Summary of Other Sample Models:

This section provides an overview of the different models used in the project to classify emotions from facial images. The models vary in complexity, architecture, and performance. Key details about each model, including the base architecture, training parameters, and accuracy achieved, are summarized below:

Base Model	Model Type	Dataset	Loss Function	Optimizer	Learning Rate	Weight Decay	Epochs	Accuracy (%)
netCNN	Custom CNN - Baseline	FER-2013	CrossEntropyLoss	SGD	0.01	-	30	57
emoCNN1	Custom CNN	FER-2013	CrossEntropyLoss	SGD	0.01	-	30	59
emoCNN2	Custom CNN	FER-2013	CrossEntropyLoss	SGD	0.01	-	30	63
VGG16	Pre-Trained CNN	FER-2013	CrossEntropyLoss	SGD	0.01	-	30	31
ResNet50	Pre-Trained CNN	FER-2013	CrossEntropyLoss	SGD	0.01	-	30	61
ResNet50	Pre-Trained CNN	FER-2013	CrossEntropyLoss	Adam	0.0001	1.00E-06	50	70

Table 1: Key Metrics and Performances of Sample Models

Vision Transformer (ViT) model

To improve our accuracy, we will use the pretrained ViT model. This model is an alternative to Convolutional Neural Networks (CNNs) and leverages the power of the transformer architecture. Unlike CNNs, which rely on convolutions to extract local patterns, ViT divides images into patches and processes them as a sequence of tokens, allowing it to capture global dependencies more effectively. This capability makes ViT particularly well-suited for complex visual recognition tasks where understanding relationships across the entire

image is crucial. Moreover, ViT scales effectively with large datasets, often surpassing the performance of CNNs when sufficient data is available. Given that we have over 160,000 images, ViT provides an ideal architecture to leverage this abundance and improve overall model performance.

We loaded the ViT model using the timm library and customized it for our classification task by modifying the final layer to match the number of emotion classes. We utilized the cross-entropy loss function, which is well-suited for multi-class classification tasks. For optimization, we employed the AdamW optimizer, a variant of the Adam optimizer that incorporates weight decay regularization. It helps prevent overfitting by penalizing large model weights. For hyperparameters, we chose a learning rate of 3e-5 with a weight decay of 0.1 to ensure stability during training. To handle memory constraints and simulate a larger batch size, we employed gradient accumulation, where gradients were updated every 8 steps instead of every single batch. This makes our effective batch size to be 256 (32x8). We also used mixed precision training with automatic casting to speed up computations on the GPU while preserving numerical stability. To help the model converge better, we used a learning rate schedule with a warmup phase followed by cosine decay. This allowed the model to start with a low learning rate and gradually increase it during the warmup phase before decaying.

4. IMPLEMENTATION

We evaluated the model regularly on the validation set to monitor its progress and prevent overfitting. Validation metrics like loss and accuracy were logged every 200 steps, and we saved the model's best weights whenever validation accuracy improved. The training process resulted in a steady improvement in both training and validation metrics. The best validation accuracy achieved was 74.70%, recorded during epoch 5 at step 20200. As shown in Figure 3 and 4, the validation loss decreased consistently across steps, while the accuracy increased. The training loss also decreased consistently. Due to the size of the training data, it took more than 2 hours to train the ViT model with 5 epoches.

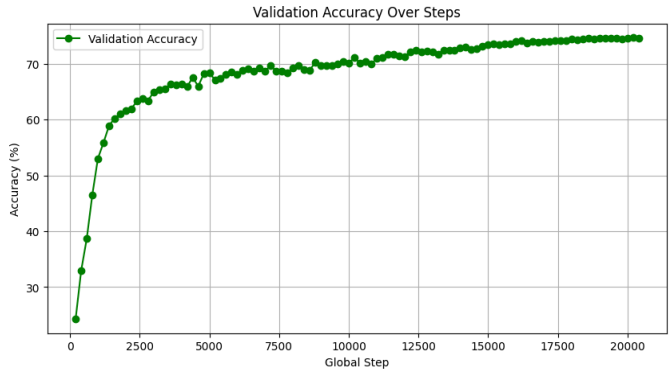

Figure 3: Validation Loss Over 20500 Steps

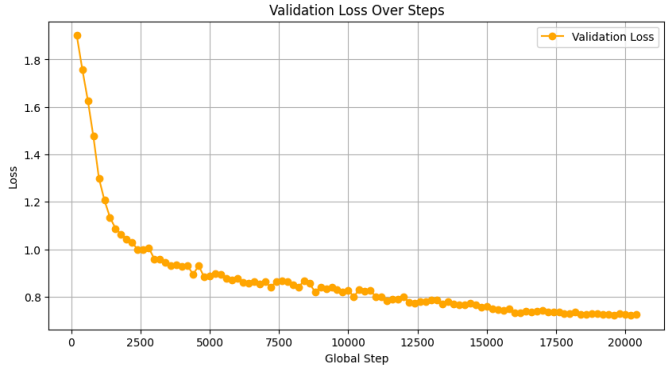

Figure 4: Validation Accuracy Over 20500 Steps

To ensure reproducibility and facilitate further use, we saved the model's weights corresponding to the best validation accuracy using PyTorch's torch.save() function.

5. RESULTS AND EVALUATION

The trained ViT model was evaluated on a test set of 33,963 images. Due to the incomparable size of images across different classes, we evaluate our model based on weighted average accuracy. Our trained ViT model achieved an overall weighted test accuracy of 72.55%. If not weighted, the accuracy is 66.62%. Below is a summary of the performance and proposed improvements we will work on.

Performance:

The model performed considerably well in classifying "neutral" (84.74% accuracy), "happy" (78.70% accuracy), indicating effective feature extraction for these distinct emotions. "Surprise" (71.23%) and "angry" (69.91%) also have relatively good accuracy. However, it has lower accuracy with "sad" (58.14%), "fear" (57.95%), and particularly "disgust" (45.66%), where high rates of misclassifications occurred.

Class	Disgust	Fear	Sad	Angry	Surprise	Happy	Neutral
Accuracy	45.66%	57.95%	58.14%	69.91%	71.23%	78.70%	84.74%

Table 2: Validation Accuracy For 7 Emotion Classes

Challenges:

The confusion matrix revealed that overlapping features between emotions contributed to misclassifications. For instance, "angry," "sad," and "fear" shared subtle facial cues, making them difficult to

distinguish, especially when expressions were less exaggerated. Similarly, “disgust” was often misclassified as “angry”. These results highlight the inherent difficulty of the task, as some expressions are ambiguous and even challenging for humans to classify accurately, which are shown in Figure 6.

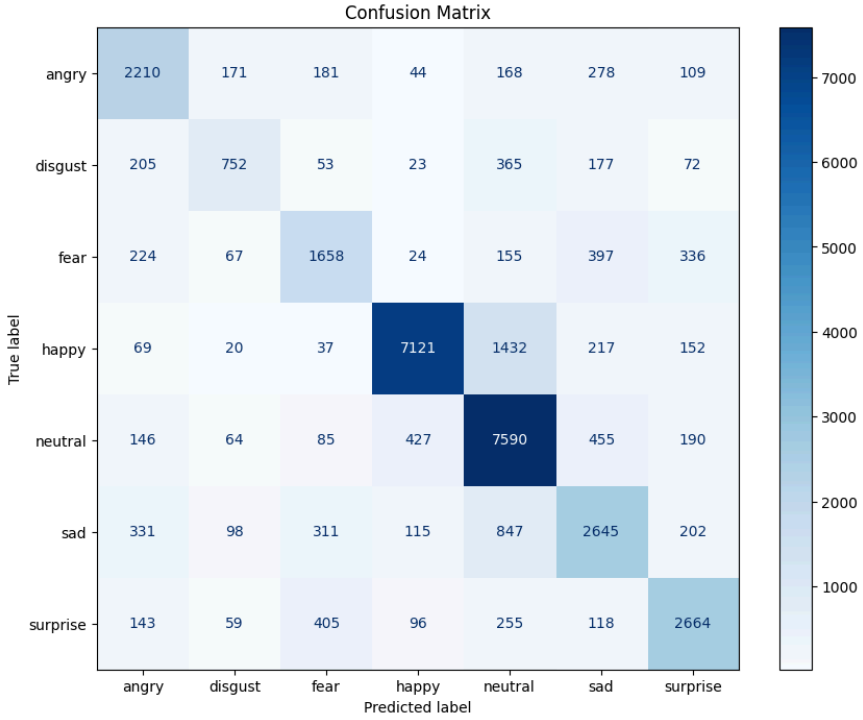

Figure 5: Confusion Matrix of Test Results

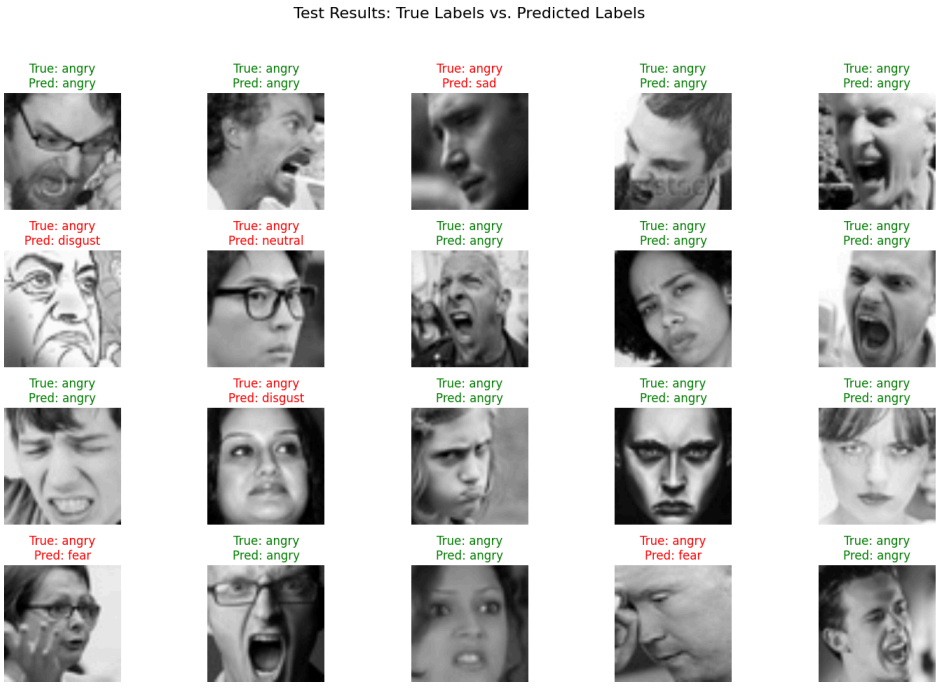

Figure 6: Sample Test Results

The dataset was imbalanced, with some classes like “disgust” and “fear” underrepresented, resulting in insufficient learning for these categories. This imbalance biased the model toward more frequent classes like “happy” and “neutral.”

Variations in lighting, pose, and occlusions (e.g., glasses, hands covering parts of the face) added further complexity, reducing the model’s robustness on certain samples.

Proposed Improvements:

To address these challenges and enhance the model's performance, we propose the following improvements:

1. We could try to utilize more advanced architectures such as pre-trained ResNet, EfficientNet, or hybrid Vision Transformer-CNN models to capture both local and global features more effectively. However, this approach may be challenging to implement given our current knowledge level in deep learning and would require significantly more time and resources to develop.
2. Finding additional data for underrepresented classes, such as “disgust” and “fear,” to address the imbalance in the dataset. Or we can try to fine-tune the pretrained ViT model with emotion-specific datasets or datasets with clearer representations of hard-to-classify emotions.
3. Fine-tune learning rate, weight decay, and other hyperparameters to optimize convergence and reduce misclassification rates.

6. DEPLOYMENT

As mentioned previously in the report, the main application we suggest for our project is in recruitment activities. In these activities, understanding facial gestures and emotional analysis can provide further insights into candidates’ engagement, stress levels, and reactions that go deeper than the answer they can input with the keyboard. For example, consulting firms can find this application very interesting given that they constantly implement aptitude tests. In these tests, the candidates go through an automated screening method that assesses them on mostly cognitive skills. By analyzing a candidate’s facial expression, the recruiter can get further

information on how the candidate deals with pressure, and oversee the emotional reactions. These reactions can show behavioral insights such as frustration or resilience which are key for certain roles.

Nevertheless, before implementing this platform, the firm should be aware of different ethical considerations. First, there are privacy concerns regarding the information of the candidate's face. Improper storage or unauthorized access to this information could lead to major privacy violations. To mitigate this, employers should make sure to ask the user for permission for recording and clearly communicate how this data will be processed and stored, and get the candidate's consent to do so. Another concern for this application is the possibility of bias. Even though our model has been trained over thousands of images and has an accuracy of (X), there is still room for error and bias. Human faces and readability of emotions vary from person to person and from race to race, and the model can lead to involuntary unfair assessments outcomes. In addition, facial expressions do not always correlate with emotions, for example, a smile might not always mean happiness. To mitigate the bias, the model should be trained further with more images, and with images with people from different demographics.

To ensure that the model leads to the lowest error, we recommend periodical audits to test the system for bias or disparities across different demographics. In addition, we recommend using this system as a supplement tool, not as a determinant to make a decision.

Lastly, another consideration is the user experience. Job applications assessments are already perceived as intimidating, and having the assessment record your face while taking it might add more pressure to the candidate.