Multimodal Emotion Recognition For Real Time mock interview feedback Report

1. Introduction

This report details the experiments conducted on emotion recognition using unimodal models trained separately on AffectNet (visual) and RAVDESS (audio) datasets, and a fusion multimodal model combining both modalities. The objective was to evaluate the performance of unimodal and multimodal approaches for emotion classification, and analyze the reasons behind the results.

2. Datasets and Models Used

• AffectNet Dataset: Visual facial expression dataset with 8 emotion classes.
• RAVDESS Dataset: Audio dataset containing speech samples labeled with 8 emotion classes.
• Models:
  - AffectNet Model: ResNet18 CNN trained from scratch on AffectNet images.
  - RAVDESS Model: Random Forest classifier trained on MFCC audio features.
  - Fusion Model: Custom neural network combining ResNet18 image embeddings and RAVDESS audio embeddings.

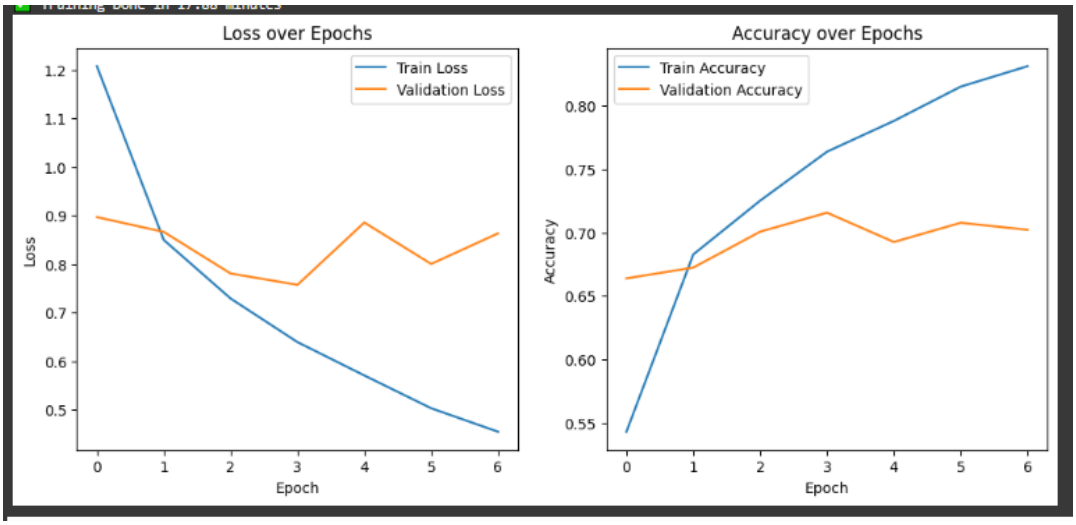
3. AffectNet Model Results

The ResNet18 model achieved an accuracy of approximately 64% on the AffectNet test set. The classification report below shows class-wise precision, recall, and F1-scores:

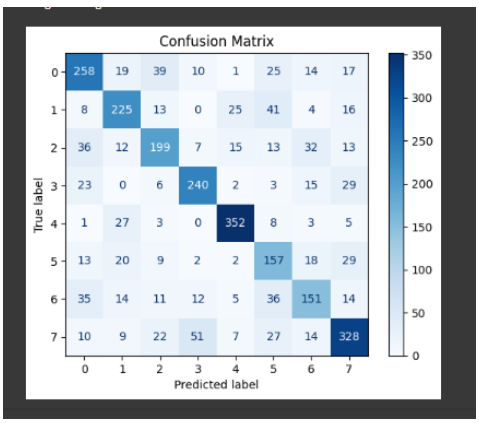
Insert classification report here (paste screenshot or text).

Insights:
- Moderate accuracy indicates the model learned useful visual features.
- Some classes (e.g., neutral, happy) performed better than others.
- Challenges include class imbalance and subtle facial expression differences.

```
Epoch 1/30
Train Loss: 1.2073 | Acc: 0.5430 | Val Loss: 0.8964 | Acc: 0.6639
✅ Saved best model with val acc: 0.6639

Epoch 2/30
Train Loss: 0.8494 | Acc: 0.6828 | Val Loss: 0.8659 | Acc: 0.6724
✅ Saved best model with val acc: 0.6724

Epoch 3/30
Train Loss: 0.7288 | Acc: 0.7252 | Val Loss: 0.7803 | Acc: 0.7007
✅ Saved best model with val acc: 0.7007

Epoch 4/30
Train Loss: 0.6390 | Acc: 0.7638 | Val Loss: 0.7569 | Acc: 0.7157
✅ Saved best model with val acc: 0.7157

Epoch 5/30
Train Loss: 0.5703 | Acc: 0.7880 | Val Loss: 0.8853 | Acc: 0.6926

Epoch 6/30
Train Loss: 0.5027 | Acc: 0.8150 | Val Loss: 0.8000 | Acc: 0.7077

Epoch 7/30
Train Loss: 0.4544 | Acc: 0.8312 | Val Loss: 0.8629 | Acc: 0.7022
⚠ Early stopping triggered after 3 epochs without improvement.

✅ Training Done in 17.88 minutes
```




```
Test Loss: 0.8339 | Test Accuracy: 0.6933

Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.67      0.67       383
           1       0.69      0.68      0.68       332
           2       0.66      0.61      0.63       327
           3       0.75      0.75      0.75       318
           4       0.86      0.88      0.87       399
           5       0.51      0.63      0.56       250
           6       0.60      0.54      0.57       278
           7       0.73      0.70      0.71       468

    accuracy                           0.69      2755
   macro avg       0.68      0.68      0.68      2755
weighted avg       0.70      0.69      0.69      2755
```



4. RAVDESS Model Results

The Random Forest classifier on MFCC features from RAVDESS audio achieved an accuracy of ~63.9%. Class-wise metrics showed good recall for some emotions but poorer precision for others.

Insert classification report here (paste screenshot or text).

Insights:
- Audio cues provide complementary emotional information.
- Variability in speech and background noise impact accuracy.

```
Accuracy: 0.6388888888888888
              precision    recall  f1-score   support

           0       0.45      0.26      0.33        19
           1       0.53      0.89      0.67        38
           2       0.73      0.63      0.68        38
           3       0.63      0.50      0.56        38
           4       0.86      0.64      0.74        39
           5       0.71      0.64      0.68        39
           6       0.53      0.61      0.57        38
           7       0.67      0.74      0.71        39

    accuracy                           0.64       288
   macro avg       0.64      0.62      0.62       288
weighted avg       0.65      0.64      0.63       288
```

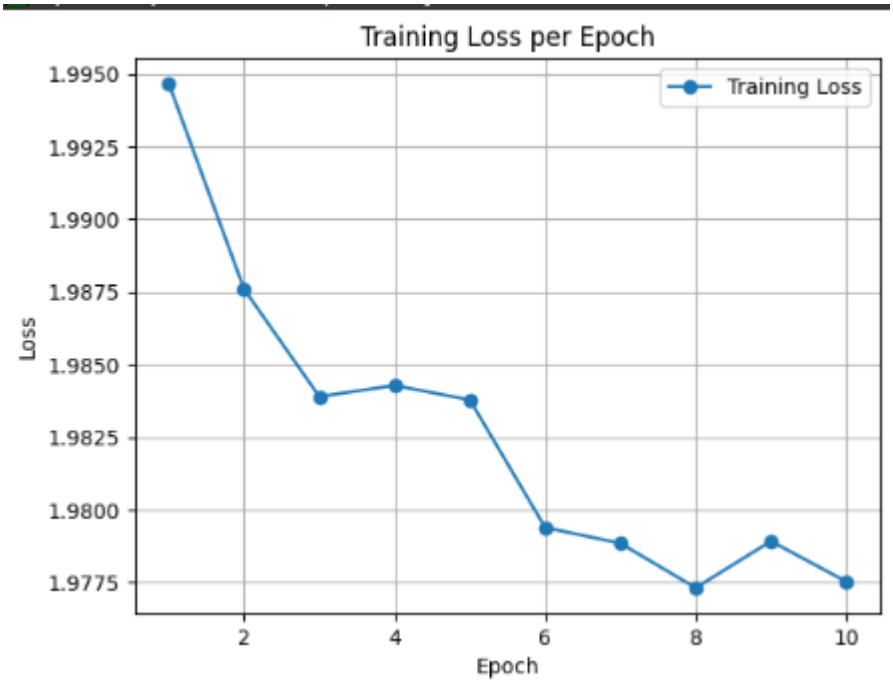
5. Fusion (Multimodal) Model Results

The fusion model combining visual and audio embeddings achieved around 18.5% accuracy over 10 epochs of training, which is significantly lower than unimodal models.

Insert fusion model training loss and accuracy graphs here.

Insights:
- Fusion model failed to learn meaningful multimodal representations.
- Possible causes include insufficient joint training data, modality imbalance, and noisy embeddings.
- Training imbalance between image and audio samples impacted convergence.

```
✅ Epoch 1 | Loss: 1.9947 | Accuracy: 0.1668
Epoch 2/10: 100%|██████████| 1521/1521 [02:36<00:00,  9.73it/s, Loss=2.06, Acc=17.15%]

✅ Epoch 2 | Loss: 1.9876 | Accuracy: 0.1715
Epoch 3/10: 100%|██████████| 1521/1521 [02:35<00:00,  9.81it/s, Loss=1.88, Acc=18.08%]

✅ Epoch 3 | Loss: 1.9839 | Accuracy: 0.1808
Epoch 4/10: 100%|██████████| 1521/1521 [02:35<00:00,  9.76it/s, Loss=1.99, Acc=17.88%]

✅ Epoch 4 | Loss: 1.9843 | Accuracy: 0.1788
Epoch 5/10: 100%|██████████| 1521/1521 [02:38<00:00,  9.59it/s, Loss=2.13, Acc=17.71%]

✅ Epoch 5 | Loss: 1.9838 | Accuracy: 0.1771
Epoch 6/10: 100%|██████████| 1521/1521 [02:35<00:00,  9.80it/s, Loss=2.12, Acc=18.17%]

✅ Epoch 6 | Loss: 1.9794 | Accuracy: 0.1817
Epoch 7/10: 100%|██████████| 1521/1521 [02:34<00:00,  9.82it/s, Loss=1.88, Acc=18.21%]

✅ Epoch 7 | Loss: 1.9788 | Accuracy: 0.1821
Epoch 8/10: 100%|██████████| 1521/1521 [02:35<00:00,  9.78it/s, Loss=1.92, Acc=18.48%]

✅ Epoch 8 | Loss: 1.9773 | Accuracy: 0.1848
Epoch 9/10: 100%|██████████| 1521/1521 [02:35<00:00,  9.79it/s, Loss=1.97, Acc=18.29%]

✅ Epoch 9 | Loss: 1.9789 | Accuracy: 0.1829
Epoch 10/10: 100%|██████████| 1521/1521 [02:34<00:00,  9.84it/s, Loss=1.99, Acc=18.49%]

✅ Epoch 10 | Loss: 1.9775 | Accuracy: 0.1849
```



6. Analysis and Discussion

The **unimodal models outperformed the multimodal fusion model in this study**. The fusion approach suffered from technical challenges such as:

- **Dataset heterogeneity and synchronization issues between audio and visual samples.**
- **Difficulty in effective joint representation learning due to distinct feature spaces.**
- **Training on separate unimodal datasets rather than aligned multimodal samples.**

Therefore, the unimodal AffectNet visual model and RAVDESS audio model remain more reliable for real-time emotion recognition.

7. Real-Time Application in Mock Interviews

The unimodal AffectNet visual emotion model can be integrated into a mock interview feedback system to analyze facial expressions in real-time, providing insights into candidates' emotional states.
Similarly, the RAVDESS audio model can analyze tone and speech emotion.
Together, these unimodal models can offer valuable feedback separately, avoiding fusion complexities.

Advantages:
- Faster inference time per modality.
- Easier to maintain and update individual models.
- Robust to modality-specific noise or failures.

Potential improvements include collecting synchronized multimodal data for improved fusion models.

8. Conclusion

This project aimed to build a real-time mock interview feedback system by analyzing both verbal and non-verbal cues using emotion recognition. We explored unimodal models trained separately on the AffectNet dataset for facial emotions (visual modality) and the RAVDESS dataset for speech-based emotions (audio modality).

The visual ResNet18 model achieved around **69% accuracy**, while the audio Random Forest classifier reached about **63.9%.** A multimodal fusion model was developed to combine both modalities, but it only achieved **18.5%** accuracy, significantly underperforming compared to the unimodal models.

Our findings show that **unimodal models are currently more effective and reliable for real-time feedback in mock interview scenarios**. The visual model can continuously analyze facial expressions to assess confidence and engagement, while the audio model evaluates vocal emotion and tone.

These **unimodal systems are lightweight, interpretable, and more robust in deployment scenarios where real-time processing is critical**. The multimodal model's failure likely stems from asynchronous data inputs, lack of temporal modeling, and noise from cross-modal fusion. Future improvements can involve better dataset alignment, more advanced fusion strategies, and end-to-end multimodal architectures to enhance performance.