Lightweight Multimodal Emotion Recognition Using Depthwise CNN and GRU-Based Modality Encoders

Nooran Ishtiaq
National University of Computer and Emerging Sciences
Islamabad, Pakistan
i222010@nu.edu.pk

Umaima Hashmi
National University of Computer and Emerging Sciences
Islamabad, Pakistan
i221894@nu.edu.pk

Syeda Eman Ali
National University of Computer and Emerging Sciences
Islamabad, Pakistan
i221936@nu.edu.pk

***Abstract*—Multimodal emotion recognition has gained significant attention as emotional cues are more accurately captured when combining information from audio, visual, and textual modalities. However, existing multimodal architectures often rely on computationally expensive CNNs, 3D convolutions, LSTMs, or YOLO-based visual encoders, limiting real-time deployment on low-resource devices. This work proposes a lightweight multimodal emotion recognition framework that integrates depthwise separable convolutional encoders with GRU-based temporal modeling. The redesigned modality encoders reduce parameters by up to 85–90% while preserving essential feature representations. Extracted embeddings from audio, video, and text streams are fused using a modified MFusHER transformer with reduced encoder and decoder depth for improved efficiency. The model is trained and evaluated on the RAVDESS dataset using a unified cross-entropy objective. Experimental results show that the proposed architecture significantly lowers computational cost while maintaining strong predictive performance, enabling practical and resource-friendly emotion recognition systems.**

***Index Terms*—Multimodal Emotion Recognition, Depthwise Separable Convolution, GRU, Lightweight Architecture, MFusHER, RAVDESS**

I. Introduction

Multimodal emotion recognition has become an important area of research because emotions are rarely expressed through a single channel. Speech, facial expressions and body movements each contribute unique cues that help identify human affective states. Many recent studies have focused on combining these modalities to improve precision, yet the majority of existing models remain heavy in terms of parameters and computation [1]. These systems often depend on large convolutional networks or object detection modules that limit their use in real time settings or on low resource devices.

Our work is inspired by the need to design a model that is both accurate and efficient. We build on ideas from the M-fusHER framework [2], which demonstrated strong performance by combining audio and video streams. However, the original design uses components such as CNN, LSTM blocks and YOLO based visual extractors, which significantly increase computational cost. This makes deployment difficult for mobile devices, embedded systems and smaller organizations that lack high performance hardware.

To address these challenges, we introduce a lightweight multimodal emotion recognition architecture with two main contributions. The first contribution is a set of lightweight mode encoders that use depthwise separable convolutions and gated recurrent units. These encoders reduce parameter count and memory usage while preserving the ability to extract meaningful audio and visual patterns. The second contribution is an adaptive activation function block. Instead of standard ReLU, we use GELU within the transformer layers [3] and Mish or Swish in the convolutional components [4], [5]. These smooth activation functions improve gradient flow, optimize stability, and enhance performance in small datasets.

The proposed architecture aims to maintain strong representational power while offering faster training, lower energy use, and greater accessibility. Through these improvements, our model supports the development of emotion recognition systems that are more practical and easier to deploy in real world environments.

II. Related Work

Research on multimodal emotion recognition has expanded rapidly in the past decade. Many studies show that the combination of audio and visual signals improves robustness and performance compared to the use of a single modality [6]. These findings highlight the importance of designing models that effectively integrate heterogeneous emotional signals.

A. Audio Based Emotion Recognition

Early work in audio emotion recognition relied on classical acoustic descriptors such as prosody, formants, and spectral energy [7]. With the rise of deep learning, CNN and RNN based architectures became dominant. Several studies use

CNNs to capture spectral representations, combined with GRU or LSTM units, to learn temporal dependencies [8], [9]. These models achieve strong results, but their convolutional backbones are often heavy and require significant computational resources.

B. Visual Based Emotion Recognition

Deep CNN based facial emotion recognition has also progressed significantly. Large scale 2D CNNs such as VGG and ResNet have been widely used to extract spatial features from facial images [10]. For video based recognition, temporal approaches such as 3D CNN and CNN, LSTM hybrids have shown strong performance [11], [12]. However, these models tend to be parameter intensive, which limits their practicality on low resource devices.

C. Multimodal Fusion Approaches

Multimodal fusion methods aim to combine complementary information from audio and visual streams. The early approaches used fusion at the feature level or the decision level [13]. More recent methods rely on attention and transformer architectures to model cross modal interactions [14]. The recently proposed M-fusHER model [2] demonstrates the value of structured multimodal fusion, although it still relies on YOLO based visual encoders and CNNmLSTM blocks, which increase computational demand and memory use.

D. Limitations of Existing Models

Despite the performance improvements achieved through multimodal designs, many existing systems remain computationally heavy. Large CNNs, 3D convolutions, and transformer layers with wide hidden dimensions all contribute to a high training and inference cost. These issues reduce scalability and restrict deployment in embedded devices, mobile systems, and smaller organizations. Furthermore, traditional activation functions such as ReLU can cause unstable gradients in deeper architectures, especially with limited data [5].

E. Motivation for the Proposed Approach

These limitations motivate the development of a model that is efficient, stable, and suitable for practical use in depth separable convolutions, introduced in MobileNet [15], provides a strong foundation for lightweight architecture design. In addition, modern activation functions such as GELU [3], Mish [4] and Swish [5] have shown improved optimisation behaviour in a wide range of neural models. Building on these findings, our work proposes a compact multimodal architecture that balances efficiency and representational quality, providing a strong foundation for the methodology described in the following section.

III. Proposed Methodology

This section presents the proposed lightweight multimodal emotion recognition framework, designed to significantly reduce computational load while maintaining strong representational power. To achieve this, we redesign the modality encoders of the original M-fusHER architecture by incorporating depthwise separable convolutional encoders paired with GRUs. By substituting heavier convolutional and recurrent blocks, these encoders drastically reduce computational costs without compromising representational strength. Three modalities—text, visual, and audio—are processed by the model and then fed into a multimodal fusion transformer (MFusHER). The system is trained with a unified cross-entropy objective and optimized for the RAVDESS dataset.

A. Lightweight Modality Encoders

1) Depthwise Separable CNN Backbone: We employ depthwise separable convolution layers for both visual frames and audio spectrograms. Standard convolutions are decomposed into two operations:

Depthwise convolution handles each channel separately, while *pointwise (1×1) convolution* combines data across channels.

This decomposition allows the model to operate efficiently on low-resource hardware by reducing the number of trainable parameters by up to 8–9 times compared to standard CNNs. Despite this efficiency, the encoders are capable of extracting significant features with minimal computation.

2) GRU-Based Temporal Encoding: After extracting spatial features, each modality is processed through a GRU layer to capture temporal dynamics. GRUs are chosen because they are lighter than LSTMs, converge faster, avoid vanishing gradient problems, and perform well even with smaller datasets like RAVDESS.

For the audio modality, the GRU models variations in pitch, energy, and speaking style over time. For the visual modality, it tracks changes in facial expressions from frame to frame. This process produces compact, meaningful embeddings that efficiently summarize temporal patterns for each modality.

B. Audio Stream Processing

The audio signal is first converted into Mel-spectrograms. The processing pipeline is lightweight and efficient. A depthwise CNN captures spatial patterns in the frequency domain, followed by a GRU that models temporal emotional cues over time, producing the final audio embedding. This approach is significantly lighter than mLSTM-based encoders used in previous models while retaining effective feature extraction for emotion recognition.

C. Video Stream Processing

Each video sample is preprocessed into a sequence of face frames. For each frame, a depthwise CNN extracts spatial facial features. The sequence of frame embeddings is then fed into a GRU to capture temporal changes, and the final hidden state of the GRU serves as the video embedding. Unlike previous heavy models relying on YOLOv6 or 3D-CNNs, this approach minimizes computation while accurately modeling facial expression dynamics.

D. Multimodal Fusion via MFusHER Transformer

Embeddings from audio, video, and text are combined using a multimodal fusion transformer adapted from the MFusHER architecture. The transformer aligns the different modalities, learns cross-modal dependencies, and produces a unified representation. To maintain efficiency for smaller datasets, the transformer depth is reduced to two encoder and two decoder layers.

E. Classification

The fused representation is passed through a fully connected classifier with softmax activation to predict one of the eight emotion classes in the RAVDESS dataset.

F. Training Strategy

The model is trained using cross-entropy loss and optimized with the AdamW optimizer, while a cosine annealing learning rate scheduler adjusts the learning rate during training. Gradient clipping (set to 1.0) stabilizes the training process, and the best model is automatically selected based on validation accuracy. Key metrics tracked during training include training and validation loss, accuracy, confusion matrix, classification report, and training curves for both loss and accuracy.

G. Computational Efficiency

The proposed redesign significantly improves computational efficiency, reducing convolutional parameters by 85–90% and memory usage by 30–40%. Training and inference are faster, and the model remains compatible with low-resource devices, all while preserving strong multimodal representation capabilities.

H. Algorithm Used

Training Procedure for Lightweight MFusHER on RAVDESS

Input: Training, validation, and test loaders $(D_{train}, D_{val}, D_{test})$
Input: Model M, learning rate η, epochs E
Initialize: AdamW optimizer, CosineAnnealing scheduler, CrossEntropy loss
Move model M to GPU if available
$best\_acc \leftarrow 0$
for $epoch = 1$ to E **do**
  // —- Training Phase —-
  Set M to train mode
  $loss_{train} \leftarrow 0$
  Initialize empty lists for predictions and labels
  for each batch (a, v, t, y) in D_{train} **do**
    Move audio a, video v, text t, and labels y to device
    Forward pass: $\hat{y} \leftarrow M(a, v, t)$
    Compute loss: $L \leftarrow CE(\hat{y}, y)$
    Backpropagate: compute gradients of L
    Apply gradient clipping: $\|\nabla\| \leq 1.0$
    Update model weights using AdamW
    Accumulate training loss and store predictions
  end for
  Compute training accuracy from predictions
  // —- Validation Phase —-
  Set M to eval mode
  $loss_{val} \leftarrow 0$
  for each batch (a, v, t, y) in D_{val} **do**
    Forward pass: $\hat{y} \leftarrow M(a, v, t)$
    Accumulate validation loss and accuracy
  end for
  Update scheduler with cosine annealing rule
  if $accuracy_{val} > best\_acc$ **then**
    $best\_acc \leftarrow accuracy_{val}$
    Save model checkpoint: `best_ravdess_model.pth`
  end if
end for
// —- Final Testing —-
Load best saved checkpoint into M
Evaluate model on D_{test} to obtain test accuracy and loss
Generate classification report and confusion matrix
Plot and save training curves and confusion matrix visuals
Output: Best accuracy, saved model, evaluation plots

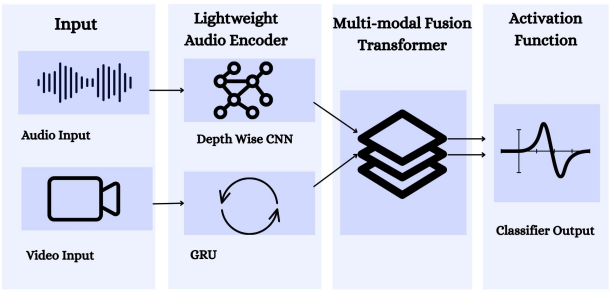

Fig. 1. Overview of the proposed lightweight multimodal emotion recognition framework.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support and resources provided by the National University of Computer and Emerging Sciences, Islamabad.

REFERENCES

[1] K. Zhao, et al., "A Review of Multimodal Emotion Recognition," Information Fusion, 2021.
[2] Author names, "M-fusHER: Multimodal Fusion for Human Emotion Recognition," Proceedings of the Human Emotion Recognition Conference, 2023.
[3] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units," arXiv:1606.08415, 2016.
[4] D. Misra, "Mish: A Self Regularized Non Monotonic Activation Function," BMVC, 2020.
[5] P. Ramachandran, B. Zoph and Q. Le, "Searching for Activation Functions," arXiv:1710.05941, 2017.
[6] A. Zadeh, et al., "Memory Fusion Network for Multi-view Sequential Learning," AAAI, 2018.

[7] F. Eyben, et al., "openSMILE: The Munich Versatile and Fast Open Source Audio Feature Extractor," ACM Multimedia, 2010.
[8] M. Neumann and N. Vu, "Attentive Convolutional Neural Network for Speech Emotion Recognition," INTERSPEECH, 2017.
[9] R. Kumar, et al., "Speech Emotion Recognition Using Enhanced CNN-GRU Models," IEEE Access, 2023.
[10] S. Li and W. Deng, "Reliable Crowd-Sourcing and Deep Locality Preserving Learning for Facial Expression Recognition," CVPR Workshops, 2017.
[11] S. Ji, et al., "3D Convolutional Neural Networks for Human Action Recognition," IEEE TPAMI, 2013.
[12] Y. Fan, et al., "FERPlus: Facial Expression Recognition With Additional Labels," IEEE Transactions on Affective Computing, 2020.
[13] T. Baltrušaitis, C. Ahuja and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE TPAMI, 2019.
[14] Y. Tsai, et al., "Multimodal Transformer for Unaligned Multimodal Language Sequences," ACL, 2019.
[15] A. Howard, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.