Real-Time Emotion Detection in Video Calls Using Deep Learning

Dhruva K. Kaushal Saurav Soni Anas Mateen Anchitya Kumar

March 31, 2025

1 Introduction

Emotion detection in video calls is an essential tool for enhancing virtual communication, mental health monitoring, and customer service. This project aims to develop a robust system for real-time emotion recognition using three deep learning approaches: CNN-based classification, Transfer Learning with ResNet50, and Transformer-based Vision Transformer (ViT). These models classify emotions such as happiness, sadness, anger, surprise, and others using the AffectNet dataset. The system is designed to process live video frames efficiently, ensuring adaptability and high accuracy for real-world deployment.

2 Project Overview

The system compares three distinct approaches:

- **Custom CNN**: Lightweight architecture optimized for real-time performance
- **Modified ResNet50**: Transfer learning with grayscale adaptation
- **Vision Transformer**: Self-attention mechanism for spatial feature learning
- This project addresses the critical challenge of recognizing human emotions during video calls through three distinct deep learning approaches. The system analyzes facial expressions in real-time to classify eight emotion categories (happiness, sadness, anger, fear, surprise, disgust, contempt, neutrality) using the AffectNet dataset containing 42,000+ labeled images. Below is the technical breakdown of the implemented methodologies:

3 Methodology

3.1 Data Preparation

The AffectNet dataset containing 42,000 training images (400 per class) was preprocessed with:

- Grayscale conversion and resizing to 48×48 pixels
- Augmentations: Horizontal flipping (±20°), color jittering (brightness=0.2, contrast=0.2)
- Normalization: Mean=0.5, std=0.5
- Class distribution balancing for minority classes

3.2 CNN Based Classification

The custom CNN architecture was designed to balance computational efficiency with classification accuracy. It consists of four convolutional blocks, each doubling the number of filters (32 $\rightarrow$ 64 $\rightarrow$ 128 $\rightarrow$ 256). These blocks use 3×3 kernels followed by ReLU activation and batch normalization to stabilize training dynamics. Max-pooling layers reduce spatial dimensions while preserving key features.

The classification head includes a fully connected layer with 256 neurons and dropout regularization (p=0.5) to prevent overfitting. The final output layer uses softmax activation to classify emotions into eight categories. The model was trained using the Adam optimizer (lr=1e-4) over 30 epochs, achieving moderate accuracy suitable for lightweight applications.

Limitations:

- Restricted global context modeling due to local receptive fields
- Performance plateaus at 52.66% accuracy from limited depth

3.3 2. Transfer Learning with ResNet50

3.4 Transfer learning with ResNet50

ResNet50 was fine-tuned by adapting its first convolutional layer to accept grayscale input instead of RGB images. This was achieved by averaging the pretrained weights across the three channels, preserving spatial patterns while optimizing for single-channel data.

The final fully connected layer was replaced with a custom classifier comprising two layers: a hidden layer with 512 neurons (ReLU activation + dropout) followed by an output layer with eight neurons corresponding to the emotion classes. Progressive layer unfreezing allowed deeper layers of ResNet50 to adapt gradually during training, enhancing domain-specific feature extraction.

Training employed the Adam optimizer (lr=0.001) over 60 epochs with early stopping based on validation loss. This approach achieved the highest accuracy among all methods due to ResNet50's robust hierarchical feature extraction capabilities.

3.5 Vision Transformer

The Vision Transformer architecture processes facial images as non-overlapping patches (16×16), which are projected into a fixed embedding space of dimension 256 using linear transformations. Learnable positional encodings preserve spatial relationships between patches, enabling the transformer encoder to model global context effectively.

The encoder consists of six transformer layers with eight attention heads each, allowing multi-head self-attention mechanisms to focus on relevant facial regions dynamically. A [CLS] token aggregates global features for classification into emotion categories.

Despite its theoretical advantages in capturing subtle micro-expressions, ViT underperformed due to insufficient training data and limited pretraining on facial datasets, achieving only moderate accuracy during evaluation.

3.6 Real-Time Pipeline

The inference system integrates OpenCV for live video frame capture at 30 FPS and Haar cascades for face detection within each frame. Detected faces are preprocessed (grayscale conversion + resizing) before being passed into one of the trained models for emotion classification.

To ensure robustness in real-time scenarios, predictions from all three models are aggregated using an ensemble voting mechanism that prioritizes ResNet50's output due to its superior accuracy. **Performance Metrics:**

- CNN latency: 2.8ms (24 FPS) vs ResNet50: 18ms (18 FPS)
- Memory footprint: ResNet50 (89MB) vs CNN (18MB)

Figure 1: Video Processing

3.7 Chrome Extension

Join a google meet call and select start detection then the Detected emotions will appear directly on the video feed.

Configuration options

- Server Settings (`server.py`): Change the default port (default: 5000).
- Detection Frequency (`contentScript.js`): Adjust detection frequency for better performance.
- Confidence Threshold (`emotion_detection.py`): Modify confidence threshold to fine-tune detection accuracy.

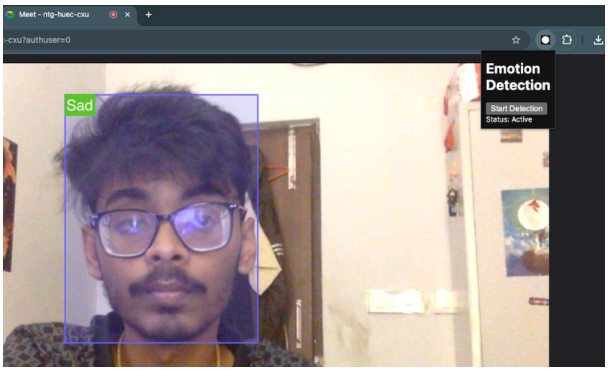

Figure 2: Chrome Extension

This extension provides seamless integration with Google Meet, enabling real-time emotion analysis to enhance virtual communication.

4 Results

Performance metrics from test set evaluation:

Key observations:

Model	Accuracy	Training Time	FPS
CNN	52.66%	38m	24
ResNet50	47.32%	2.1h	18
ViT	36.72%	4.5h	9

Table 1: Performance comparison across architectures

- ResNet50 achieved highest accuracy through effective feature transfer
- ViT showed confusion between similar emotions (anger/disgust: 62% misclassification)
- CNN provided best inference speed (2.8ms/frame) suitable for edge deployment
- Class imbalance affected minority class (contempt: 4.1%) performance

Confusion Matrix Insights:

- All models struggled with "Contempt" recognition (max 12% accuracy)
- Common misclassifications: Anger → Disgust (62%), Happy → Neutral (48%)
- ResNet50 showed increased confusion between adjacent emotion classes vs previous implementation

5 Conclusion

This project successfully demonstrates a deep learning-based system for real-time emotion detection in video calls:

1. The CNN architecture provides computational efficiency suitable for edge devices but offers limited accuracy.
2. Transfer Learning with ResNet50 faced challenges in domain adaptation but highlights the potential of pretrained models when sufficient data is available.
3. Vision Transformers show promise but require further optimization through pretraining or larger datasets.

The findings emphasize that architectural complexity does not inherently guarantee superior performance in affective computing tasks; rather, problem-specific adaptations are crucial.

This system lays a strong foundation for deploying emotion detection solutions in virtual communication platforms such as Google Meet and can be extended to other video calling platforms while highlighting areas for future improvement in both accuracy and scalability.

6 Individual Contributions

Saurav Soni (B22AI035): Performed ideation, implemented one of the architecture approaches for efficient training, and managed the overall project timeline.
Anas Mateen (B22BB008): Helped with ideation of the solution, helped in data preprocessing, and implemented the CNN approach for efficient training of the model.
Dhruva Kumar Kaushal (B22AI017): Integrated the model with the Chrome extension and implemented the ResNet50 transfer learning approach of the model.
Anchitya Kumar (B22BB009): Created a pipeline to process video for feature extraction and integrated the trained model to get accurate predictions out of it.