

Preliminary Report: CNN-Based Facial Expression Recognition on RAF-DB with Explainable AI

Zeynep Belde Alena Daneker Cedric Fernolend

{z.belde,a.daneker,cedric.fernolend}@campus.lmu.de

Abstract

This report presents a preliminary study on facial expression recognition using convolutional neural networks (CNN) trained on the Real-world Affective Faces Database (RAF-DB). To address the interpretability challenges of deep learning models, Grad-CAM is integrated as an explainability framework, enabling verification that the network attends to meaningful facial regions. This report outlines the proposed approach. Implementation and experimental evaluation will be conducted in future work.

1. Introduction

1.1. Motivation

Facial emotion recognition is a challenging task in computer vision with applications in many Human-Computer Interactions. Automatic classification of facial expressions enables systems to respond adaptively to users. However, recognizing emotions from facial images is a challenge due to high intra-class variability and inter-class similarity. Additionally, deep learning models often operate as black boxes, making it unclear whether they rely on meaningful facial features or spurious correlations.

1.2. Objective

This project aims to develop a CNN-based system that classifies facial images into six discrete emotion categories (happiness, surprise, sadness, anger, disgust, fear) while providing interpretable explanations for its predictions. Grad-CAM will be integrated to visualize which facial regions the model attends to. By combining classification accuracy with visual explainability, the target is to build a system that is effective and trustworthy.

2. Related Work

The Real-world Affective Faces Database (RAF-DB) was introduced by Li et al. (2017) and has become a widely used benchmark for facial emotion recognition. It contains

approximately 30,000 facial images collected from the internet, labeled either as one of the seven basic emotion categories (angry, disgust, fear, happy, sad, surprise, neutral) or as a compound emotion [8]. Unlike lab-controlled datasets, RAF-DB includes variations in pose, lighting, occlusion, and diverse demographic backgrounds, making it more representative of practical applications.

Custom Lightweight CNN-based Model (CLCM) has reached 84% accuracy [5] and FARNet reports 87.65% accuracy [3] on the Dataset. These results show the potential for CNN architectures to achieve high accuracy when combined with appropriate training strategies.

3. Proposed Approach

3.1. Dataset

Given the requirement to use RGB data, *RAF-DB (Real-world Affective Faces Database)* is selected as the primary dataset [8].

The images in the RAF-DB dataset do not have a fixed resolution. They are provided as aligned and cropped RGB facial images with varying spatial dimensions.

In addition to the basic emotion annotations, RAF-DB also provides labels for compound emotions, which represent combinations of two basic affective states [14]. In this project, only the basic emotion subset is used, which consists of 15,339 images annotated with seven basic emotion categories.

3.2. Preprocessing

For this project, only the basic emotion categories are required. As the compound emotion annotations are not included in the selected labeling scheme, no additional preprocessing steps are required to remove them.

Although RAF-DB includes a *neutral* emotion category, this class is not required for the objectives of this project. Consequently, all images whose majority annotation corresponds to the neutral class are excluded from the dataset. The remaining 12,815 images are used for training and evaluation, resulting in a six-class emotion classification problem.

The training process uses a soft-labeling strategy. The soft label distributions for images are renormalized to sum to one over the six target emotion categories. Accordingly, the output layer of the convolutional neural network is adjusted to match the reduced number of classes.

The faces in the RAF-DB images are already cropped and aligned; therefore, no additional face detection or alignment preprocessing is required. All images are resized to a fixed resolution of 64×64 pixels to ensure compatibility with the CNN architecture.

Furthermore, RGB color channel normalization is applied. This normalization improves training stability, accelerates convergence, and ensures a balanced contribution of all color channels during CNN training [4].

3.3. Model Architecture

The proposed model is a convolutional neural network designed for the classification of facial expression on 64×64 RGB input images, normalized in the range $[-1, 1]$. The architecture consists of three convolutional blocks followed by a classification head.

All convolutional layers employ 3×3 filters to increase the effective receptive field while maintaining a compact architecture [12]. The first two blocks each contain two convolutional layers with batch normalization [7] and ReLU activation, followed by 2×2 max pooling for spatial down-sampling. The third block extends this design to three convolutional layers to capture higher-level semantic features. Identity skip connections [6] are incorporated in each block to facilitate gradient flow during training.

For classification, global average pooling [9] reduces the feature maps to a 128-dimensional vector, followed by a fully connected layer with 64 neurons, dropout ($p=0.5$) [13], and a softmax output layer for the six emotion classes. This design prioritizes parameter efficiency and generalization.

3.4. Training

Data augmentation is applied during training, including random horizontal flips, rotations ($\pm 15^\circ$), and random translations up to 10% of the image dimensions to improve generalization. A soft label strategy is employed, utilizing the annotation distributions from RAF-DB's crowdsourced labels as targets for categorical cross-entropy loss.

The model is trained using the Adam optimizer with an initial learning rate of 3×10^{-4} and a batch size of 32 or 64, depending on available GPU resources. The official RAF-DB data split is adopted. Approximately 80% of the dataset is used for training, while the remaining 20% is reserved for testing. From the training portion, 10% of the samples are held out to form a validation set, which is employed for performance monitoring, hyperparameter tuning, and early stopping. Implementation and experimental evaluation will be carried out in future work.

3.5. Explainable AI

The inherent opacity of Convolutional Neural Networks presents a significant "black box" challenge, and relying on uninterpretable models for decision making compromises trust and conceals potential failure modes [10]. To address this, Explainable AI (XAI) frameworks are employed as essential techniques for ensuring algorithmic transparency and accountability [1].

For this project, Grad-CAM [11] was selected as the primary visualization tool. This decision is driven by the specific requirements of facial expression analysis. It is critical to verify that the model attends to valid facial features rather than spurious background noise.

While standard Grad-CAM is robust, Grad-CAM++ [2] is also considered as a refinement strategy. This improved iteration offers sharper localization capabilities, which may prove necessary if the baseline model struggles to distinguish between subtle, fine-grained facial feature configurations.

4. Evaluation Plan

Given the challenges of training from scratch on a real-world dataset, the initial target is to achieve more than 75% accuracy on the held-out RAF-DB test set. This is a realistic but challenging goal that signifies the model has learned meaningful representations.

A confusion matrix will be generated to identify which emotion pairs are most challenging for the model (e.g., Fear vs. Surprise, Disgust vs. Anger). This analysis will directly inform future improvements.

5. Timeline

Weeks 1–2 (Baseline Implementation): Set up of the data pipeline, implementation of the baseline model with soft-label loss, and conduction of initial training.

Weeks 3–4 (Analysis & Iteration): Analysis of the baseline's errors (confusion matrix, failure cases) and running of controlled experiments with architectural modifications, documenting the impact on validation performance.

Week 5 (Final Evaluation): Training of the final selected model and reporting of its definitive performance on the untouched test set.

References

- [1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable

artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. 2

- [2] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018. 2
- [3] Long Duongthang. Driver facial emotion tracking using an enhanced residual network with weighted fusion of channel and spatial attention. *Scientific Reports*, 15(1):12675, 2025. 1
- [4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 2
- [5] Mustafa Can Gursesli, Sara Lombardi, Mirko Duradoni, Leonardo Bocchi, Andrea Guazzini, and Antonio Lanata. Facial emotion recognition (fer) through custom lightweight cnn model: Performance evaluation in public datasets. *IEEE Access*, 12:45543–45559, 2024. 1
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2
- [7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 2
- [8] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. 1
- [9] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2014. 2
- [10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. 2
- [11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2
- [12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2
- [13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 2
- [14] Wei Yan, Xiaoyang Wang, Xin Li, and Guoying Zhao. Learning deep representations for compound facial expressions. *IEEE Transactions on Affective Computing*, 11(2): 317–331, 2020. 1