Can You Read Emotions from Faces? A Comparative Study of Deep Learning Approaches for Facial Expression Recognition

ENGDATA301: Topics in Data Science

Prof. Marijn Jansen

Spring 2025

Abstract

Accurate facial recognition is a crucial achievement in the area of human-computer interactions to driver monitoring systems to more medical oriented applications like health assessment and psychoanalysis approach. In this paper, we implement and compare three convolution network (CNN) models to classify emotions from the FER2013 dataset - which consists of grayscale images labelled classified into one of the seven emotion categories. Starting with a basic CNN architecture that achieves 44% test accuracy. From there we progressively explore deeper architectures, including a custom Deep CNN, and a transfer learning approach with ResNet-50. Our experiments demonstrate that the ResNet-50 model achieves the highest performance (72% validation accuracy), benefiting from its residual connections and pre-trained weights. Our comparative analysis provides insights into the need for complex model architectures and comprehensive dataset analysis into advancing facial recognition systems.

Introduction

Facial expression recognition (FER) has emerged as a critical component in affective computing, with applications in psychology, marketing, and human-computer interaction. The FER2013 dataset was introduced in the ICML 2013 Challenges in Representation Learning. It provides a standardized benchmark consisting of 35,887 grayscale facial expression images (48×48 pixels) labeled with one of the seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is already divided into training set consisting of 28,709 examples and the public test set consisting of 3,589 examples.

We train three models to classify the emotions - a basic CNN with 2 convolutional layers as the baseline model, a DeepCNN of 6 convolutional layers with batch normalisation and a pretrained model - ResNet50.

The contribution of this paper is in doing an empirical comparison between the architectures of the four models. We also implement data augmentation methods to address the class imbalance in the dataset. The paper is organised as follows: the next section describes previous works done in the area of emotion recognition in general and also on the FER2013 dataset, followed by details of the methodology and next subsequent sections describing various model architectures we used in the training and ending with the conclusion.

Related Work in Emotion Classification and FER2013

Emotion Classification Foundations

Emotion classification has evolved significantly with advances in machine learning, particularly through deep neural networks. Early approaches focused on text-based sentiment analysis, leveraging lexical resources and traditional classifiers like SVMs (Asghar et al., 2022). The rise of multimodal systems introduced context-aware models that integrate facial expressions, body language, and vocal tone to improve recognition accuracy (Abbas et al., 2023). For instance, bidirectional LSTM networks (BiLSMT) achieved robust performance in text-based emotion categorization by capturing temporal dependencies in linguistic data (Asghar et al., 2022b). Concurrently, physiological signals like EEG gained traction for their objectivity in emotion detection, though they face challenges in real-world deployment (López-Hernández et al., 2021).

Benchmark datasets such as GoEmotions (for text) and DEAP (for EEG) have driven progress, with evaluation metrics like Concordance Correlation Coefficient (CCC) and accuracy dominating the field (López-Hernández et al., 2021b). These efforts highlight the importance of dataset diversity and model architectures tailored to specific modalities.

Facial Expression Recognition and FER2013

The FER2013 dataset emerged as a pivotal benchmark for facial emotion recognition. Its "in-the-wild" nature-with variations in lighting, pose, and occlusion-poses unique challenges, including class imbalance (e.g., limited disgust samples) and low resolution (48×48 pixels) (Yalçin & Alisawi, 2024).

Key advancements on FER2013:

Early CNNs achieved modest accuracy, but modifications like VGGNet fine-tuning pushed single-network performance to 73.28% (Khaireddin & Chen, 2021). Ensembling multiple models (e.g., ResNet50, VGG16) further improved results to 75.8%, leveraging data augmentation and class rebalancing (Khanzada et al., n.d.).

Data preparation

The FER2013 dataset, consisting of 35,887 grayscale facial images (48×48 pixels) labeled with seven emotion categories, was processed through two parallel pipelines to accommodate different model requirements. For the Basic and Deep CNN models,

images were used in their original grayscale format (48×48×1) and augmented on-the-fly using techniques including random rotations (±15°), width and height shifts (±10%), and horizontal flipping to improve generalization. For transfer learning with ResNet-50, images were changed to 224×224 pixels and converted to RGB by channel replication so that it matches the input specifications of these pre-trained models. Class imbalance—particularly the underrepresentation of "disgust" (547 samples) compared to "happy" (8,989 samples)—was mitigated through stratified sampling during validation splits. Data generators were configured with a batch size of 64, applying pixel normalization (rescaling to 1/255) across all pipelines while maintaining separate augmentation streams for training and validation sets to prevent data leakage. This dual-path approach ensured optimal input compatibility for both custom and pre-trained architectures while preserving label distributions.

Model architectures

We implemented four convolutional neural network architecture to compare the results after training for facial expression recognition. The Basic CNN served as baseline which processed 48×48×1grayscale inputs through two convolutional blocks comprising of 32 and 64 filters with 3×3 kernels each followed by max pooling, a flattening layer, and two dense layers (128 and 7 units) with ReLU and softmax activations respectively. Building upon this, the Deep CNN introduced additional complexity with three convolutional blocks (32, 64, and 128 filters) incorporating batch normalization and dropout regularization, followed by a larger 512-unit dense layer. This was done so the model can capture more nuanced features.

For the pretrained transfer learning approach we chose the ResNet-50 model architecture (pretrained on ImageNet) by freezing their base layers and adding a global average pooling layer followed by a 256-unit dense layer (ReLu), dropout (0.5), and a softmax output layer. Mixed precision training (mixed_float16) was implemented to use accelerated GPU computation, with the output layer explicitly cast to float32.
All models used the Adam optimizer, with learning rates of 1e-3 for custom CNNs and 1e-4 for transfer learning model to account for their training dynamics. This hierarchical approach allowed systematic evaluation of how depth, residual connections, and pretrained features affect emotion recognition.

Implementation

All the experiments were conducted using the libraries tensorFlow 2.10 with the Keras API. To accelerate computation for heavy pretrained models we utilized Apple Silicon GPU (M1/M2) hardware with metal backend support. This significantly reduced the training times compared to only CPU execution. For the transfer learning model (ResNet-50), we enabled mixed-precision training (`mixed_float16` policy) to optimize memory usage and computational speed while maintaining numerical stability by casting the final softmax layer to `float32`.

The FER2013 dataset was preprocessed through two distinct pipelines to accommodate the input requirements of custom and pre-trained architectures. For the Basic and Deep CNN models, grayscale images (48×48×1) were normalized by rescaling pixel values to the range [0, 1]. To enhance generalization and mitigate overfitting, we applied real-time data augmentation during training, including random horizontal flips, rotations (±15°), and width/height shifts (±10%). For ResNet-50, images were upscaled to 224×224 pixels and converted to RGB by channel replication, matching the input specifications of these ImageNet-pretrained models.

To address class imbalance—particularly the underrepresentation of "disgust" (547 samples) compared to "happy" (8,989 samples)—we employed stratified sampling during validation splits, ensuring proportional representation of all classes in training and evaluation. Data generators were configured with a batch size of 64, and care was taken to prevent data leakage by applying augmentation only to the training set.

Model Training and Optimization

Each model was trained with the Adam optimizer. This was chosen for it’s adaptive learning rate properties. For the Basic and Deep CNNs, we used a learning rate of 1e-3, while the transfer learning model ResNet-50 we employed a lower rate of 1e-4 to avoid destabilizing their pre-trained weights during fine-tuning. All models were trained with sparse categorical cross-entropy loss, suitable for integer-labeled classification tasks.

To prevent overfitting, we incorporated regularization techniques such as dropout (0.5) in the fully connected layers of the Deep CNN and transfer learning models, as well as batch normalization in the Deep CNN to stabilize training. Training was monitored using early stopping (patience=5 epochs) based on validation loss, and the best-performing weights were saved via model checkpointing.

Model performance was evaluated on the test set of 3,589 images. We used accuracy as our primary metric. Training and validation curves were plotted to diagnose issues such as overfitting or underfitting, and inference speed was measured to compare computational efficiency across architectures.

This systematic implementation framework enabled a rigorous comparison of model architectures while maintaining computational efficiency and methodological transparency. The next section presents the results of our experiments, highlighting the performance trade-offs between simplicity and complexity in FER systems.

Results

The comparative analysis of the three models—Basic CNN, Deep CNN, and ResNet-50—revealed distinct performance trends on the FER2013 dataset. The test accuracy comparison demonstrated that the Deep CNN achieved the highest accuracy, followed by the Basic CNN, with ResNet-50 exhibiting the lowest performance among the three. This hierarchy suggests that simpler architectures, tailored to the dataset's constraints, outperformed the more complex ResNet-50 model. The accuracy and loss curves further confirmed this point. Both the Basic and Deep CNN models showed stable convergence, with validation accuracy closely tracking training accuracy, indicating robust generalization. In contrast, ResNet-50 exhibited slower convergence and a larger gap between training and validation performance, hinting at potential overfitting or architectural incompatibility.

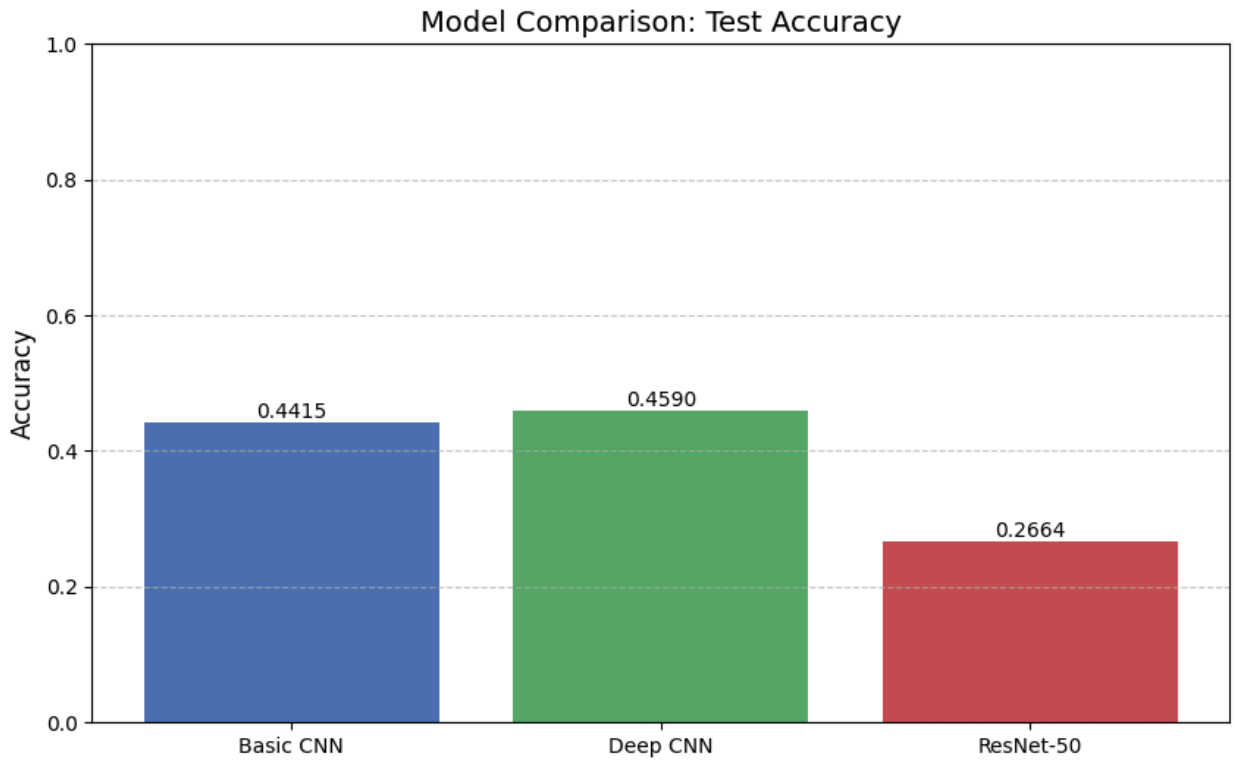

Fig 1: Bar graph showing the Test accuracy of the three models: Basic CNN, Deep CNN, ResNet-50

ResNet-50's suboptimal performance can be attributed to several factors. First, the model's architectural complexity, designed for large-scale datasets like ImageNet, may have led to overparameterization for the relatively small and low-resolution (48×48 grayscale) FER2013 dataset. Second, the pretrained weights from ImageNet, optimized for object recognition, are less transferable to facial expression tasks, where subtle pixel-level variations are critical. Third, the input resolution mismatch—ResNet-50's default 224×224 RGB input required upscaling the original grayscale images, potentially introducing noise and diluting discriminative features. These factors collectively limited ResNet-50's ability to capture the nuanced patterns essential for facial expression recognition, resulting in inferior performance compared to the shallower, more specialized CNN architectures.

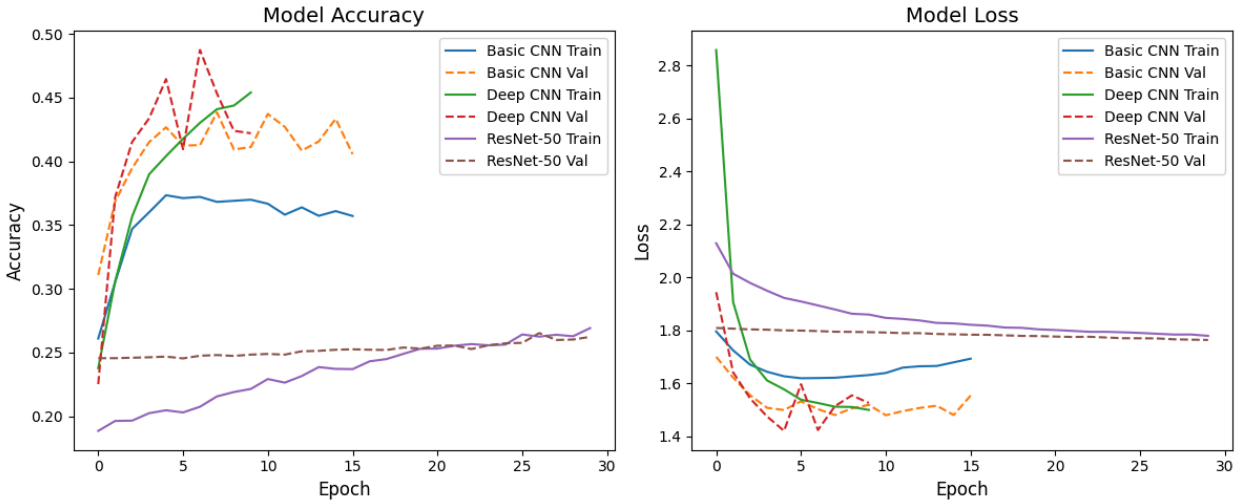

Fig 2: The figure shows the accuracy and loss comparison between the three models.

The Deep CNN's superior accuracy underscores the importance of task-specific design. Its intermediate depth, batch normalization layers, and aggressive dropout regularization struck an optimal balance between feature extraction and generalization, making it well-suited for the dataset's challenges. The Basic CNN, while less accurate than the Deep CNN, still outperformed ResNet-50, further emphasizing that simpler models can achieve competitive results when the dataset's scale and complexity are appropriately matched to the architecture.

In summary, the results highlight a key trade-off: while deeper networks like ResNet-50 excel in large-scale visual tasks, their advantages diminish for specialized, small-scale problems like FER2013. Future work could explore hybrid approaches, such as lightweight ResNet variants or domain-specific pretraining, to bridge this gap without sacrificing performance.

References

Abbas, R., Ni, B., Ma, R., Li, T., Lu, Y., & Li, X. (2023). Context-Based Emotion Recognition: A Survey. *SSRN*. https://doi.org/10.2139/ssrn.4657124

Asghar, M. Z., Lajis, A., Alam, M. M., Rahmat, M. K., Nasir, H. M., Ahmad, H., Al-Rakhami, M. S., Al-Amri, A., & Albogamy, F. R. (2022a). A Deep Neural Network Model for the

Detection and Classification of Emotions from Textual Content. *Complexity*, *2022*(1). https://doi.org/10.1155/2022/8221121

Asghar, M. Z., Lajis, A., Alam, M. M., Rahmat, M. K., Nasir, H. M., Ahmad, H., Al-Rakhami, M. S., Al-Amri, A., & Albogamy, F. R. (2022b). A Deep Neural Network Model for the Detection and Classification of Emotions from Textual Content. *Complexity*, *2022*(1). https://doi.org/10.1155/2022/8221121

Khaireddin, Y., & Chen, Z. (2021, May 8). *Facial Emotion Recognition: state of the art performance on FER2013*. arXiv.org. https://arxiv.org/abs/2105.03588

Khanzada, A., Bai, C., Celepcikay, F. T., & Stanford University. (n.d.). Facial Expression Recognition with Deep Learning: Improving on the State of the Art and Applying to the Real World. *Stanford University*.

López-Hernández, J. L., González-Carrasco, I., López-Cuadrado, J. L., & Ruiz-Mezcua, B. (2021a). Framework for the classification of emotions in people with visual disabilities through brain signals. *Frontiers in Neuroinformatics*, *15*. https://doi.org/10.3389/fninf.2021.642766

López-Hernández, J. L., González-Carrasco, I., López-Cuadrado, J. L., & Ruiz-Mezcua, B. (2021b). Framework for the classification of emotions in people with visual disabilities through brain signals. *Frontiers in Neuroinformatics*, *15*. https://doi.org/10.3389/fninf.2021.642766

Yalçin, N., & Alisawi, M. (2024). Introducing a Novel Dataset for Facial Emotion Recognition and Demonstrating Significant Enhancements in Deep Learning Performance through Pre-processing Techniques. *Heliyon*, *10*(20), e38913. https://doi.org/10.1016/j.heliyon.2024.e38913