CNN emotion detection

Jason Moore

November 2024

1 Overview

Emotion detection was chosen as a focus for this project due to its applications in enhancing human-computer interactions, sentiment analysis, and various psychological and security domains. By employing computer vision techniques, this project aimed to classify human emotions from facial expressions in real-time.

2 Introduction

This project utilizes deep learning and computer vision to analyze the facial expressions and categorize them into predefined emotion categories. Using the FER-2013 dataset the system was designed to preprocess, train, and validate the CNN model to recognize emotions effectively. The FER dataset contains grayscale images of human faces each sized 48x48 pixels. It is organized into seven emotion categories angry, happy, sad, surprised, disgust, fear, and neutral. The simplicity and standardization make it ideal for training convolutional neural networks while enabling comparisons across studies.

3 Research

The insights gathered are from *"Real Time Facial Emotion Recognition Using Deep Learning"* by Naveen N.C., Sai Smaran K.S., and Shamitha A.S. Their research demonstrated that CNNs outperformed other models like ResNEt-50, RNNs, DTSCNNS in terms of accuracy, processing, speed, and robustness. In the paper they explained that CNNs acheived a peak accuracy of 96.03% on the dataset which makes it the most effective model to experiment.

4 Methodology

The project began with data preprocessing resizing images, normalization, and their label one hot encoded. Data augmentation such as flipping and rotation were applied to improve generalization. The CNN architecture begins with an initial convolutional layer that applies filters to the input image to detect low-level features such as edges and corners. Each convolutional block is followed by a batch normalization layer to stabilize and accelerate training by normalizing the activations. Max-pooling layers are incorporated after specific convolutional layers to reduce the spatial dimensions, thereby decreasing computational complexity and focusing on the most salient features.

4.1 Key Highlights of the Architecture

- **Convolutional Layers:** The network employs several convolutional layers with increasing filter sizes, starting from 32 filters in the first layer and expanding up to 128 filters in the deeper layers. These layers capture both fine-grained details and high-level patterns in the facial images, critical for differentiating subtle emotional expressions.

- **Dropout Layers:** Dropout is applied throughout the network to prevent overfitting. For instance, dropout rates are strategically included after pooling layers and dense layers to ensure generalization across unseen data.

- **Dense and Flatten Layers:** The flatten layer converts the multi-dimensional output from the convolutional layers into a one-dimensional vector. This vector is then passed through fully connected dense layers. The dense layers have 2,048 and 1,024 units, respectively, with ReLU activation, followed by a final output layer with 7 units (representing the emotion categories) using a softmax activation function.

- **Batch Normalization:** Batch normalization layers are consistently integrated to reduce internal covariate shift, ensuring that the model trains faster and more reliably.

This architecture balances depth and computational efficiency, enabling it to achieve high accuracy while maintaining the speed necessary for real-time emotion detection. The strategic use of convolutional blocks, dropout, and dense layers ensures that the model is well-suited for recognizing nuanced patterns in facial expressions.

5 Results

The CNN model demonstrated strengths in detecting prominent emotions but faced challenges with imbalanced data and subtle emotion distinctions.

5.1 Training and Validation Performance

Training accuracy improved to 65%, while validation accuracy plateaued at 55%. Training loss decreased consistently, but fluctuations in validation loss indicate slight overfitting.

5.2 Classification Metrics

- **Best Performance:** *Happy* achieved the highest precision (0.74), recall (0.86), and F1-score (0.79), followed by *Surprise* with an F1-score of 0.70.
- **Poor Performance:** *Fear* and *Disgust* showed low recall (0.38 and 0.24), affected by class imbalances and subtle visual distinctions.

The macro-average F1-score was 0.53, reflecting uneven class performance.

5.3 Test Evaluation

The model achieved a test accuracy of 70.45% and a test loss of 0.7096, showing reasonable generalization to unseen data.

5.4 Observations

Challenges include slight overfitting, class imbalances, and difficulty with subtle emotional variations (e.g., *Fear* vs. *Surprise*). Future improvements may include advanced architectures and better dataset balancing. The model successfully identified two test images, accurately predicting the emotions ”happy” and ”surprised.” However, it misclassified the ”sad” image, likely due to the visual similarity between ”sad” and ”angry,” which can be challenging to differentiate. Despite this, the model demonstrated its effectiveness by correctly recognizing the ”happy” and ”surprised” emotions.

6 Personal Takeaways

This project demonstrated the advancements of neural networks and how they still perform strong for computer vision tasks. Despite the popularity of transformer models which may perform better for such task neural networks like CNN still offer simplicity which may lead to better performance. The model showcased its strengths and weaknesses with identifying emotions. Real time object detection is getting better as more research is conducted and this project aims to contribute to the ever growing pursuit for complex computer vision tasks.