Masters in Data Science

Facial Expression Recognition

ResNet50 vs our CNN

Akshay M(am3364), Atharva H(ah1377), Vidisha K(vk434), Vishwa P(vup4)

Abstract

Facial Expression Recognition (FER) is a specialized area of computer vision that focuses on detecting human emotions automatically by analyzing facial expressions. FER has a wide range of practical applications, such as improving human-computer interactions, enhancing virtual reality experiences, and monitoring mental health. Recently, Convolutional Neural Networks (CNNs) have exhibited exceptional results in FER tasks. Therefore, we suggest a CNN-based FER model that can effectively categorize seven fundamental emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) by examining facial images.

Problem

Recognizing human emotions is a crucial aspect of improving human-computer interactions and has numerous practical uses. With the advancement of deep learning and computer vision techniques, Facial Emotion Recognition (FER) has become a popular research domain in recent times. However, accurately recognizing human emotions from facial images still poses a significant challenge due to several factors such as lighting, facial expression, occlusion, and individual differences. Hence, there is a pressing need for robust and precise FER models that can perform effectively on different datasets and in diverse situations. Human accuracy for this dataset is 70%This project aims to develop a CNN-based FER model capable of accurately categorizing human emotions from facial images and surpassing existing FER models' performance(ResNet50 based model).

Data

The FER 2013 dataset by Ian Goodfellow is a widely used benchmark dataset for Facial Emotion Recognition (FER) tasks. It consists of 35,887 grayscale images of size 48x48 pixels. The dataset has seven emotion categories, including anger, disgust, fear, happiness, sadness, surprise, and neutral. The images are collected from various sources, including the internet and human-labeled datasets. The FER 2013 dataset is commonly used for training and evaluating FER models.

Methodology

Data Preprocessing

Preprocessed the dataset to ensure that the images are of the same size, have the same color channels, and are properly labeled. Later we handled the categorical emotion variable by one hot encoding technique.

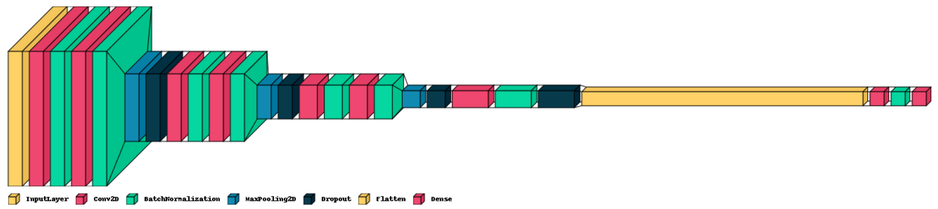
Data Augmentation

In order to improve the robustness and generalization of the model, we augmented the image data further by rotating images on range of 20 degrees, shearing, horizontally flipping, normalizing pixel range from [0-225] to [0-1] , splitting the train data in 20/80 validation training split.

Model Architecture and Training

As our project aims to develop a CNN-based FER model capable of accurately categorizing human expressions from facial images and surpassing existing FER models' performance(ResNet50 based transfer Learning model) we started with using the ResNet50 architecture as a base which is pre-trained on the ImageNet dataset, which contains millions of images from thousands of classes. Added a flatten layer to convert the output tensor of the ResNet50 model to a 1D tensor, two dense layers with 128 and 64 units respectively, and activation functions of ReLU. The final dense layer has 7 units with a softmax activation function, which is used to output the probability of each class.

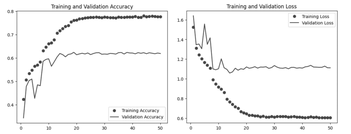
Our model consists of several convolutional layers, batch normalization, max pooling, and dropout layers. The input layer takes an image tensor of shape (48,48,3) as input, and the output layer has 7 units, one for each class in the FER task, with a softmax activation function. The model gradually increases the number of filters from 64 to 512 across convolutional layers.



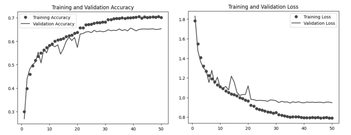
Models are trained with categorical_crossentropy as loss function, adaptive moment estimation(Adam) optimizer for 50 epochs with a batch size of 32. ‘ModelCheckpoint’ callback saves the best model based on the validation loss metric during training and ‘ReduceLROnPlateau’ callback reduces the learning rate by a factor of 0.3 if the validation loss does not improve for 3 epochs.The minimum learning rate is set to 0.000001 and starting with learning rate of 0.001.

Findings

We observed that ResNet50 model started with an higher accuracy as compared to our CNN model. This is because of the fact that ResNet is pre-trained on a large dataset (ImageNet) with millions of images from thousands of classes. This allowed the network to learn a rich set of feature representations that can be useful for many computer vision tasks, including facial expression recognition.
While on the other hand CNN model is designed specifically for Facial Expressions. Although the network architecture was carefully designed, the initial weights of the network were randomly initialized, which means that the network had to learn the features from scratch during training.



ResNet50 based model



Our CNN Model

Model	Recall	Precision	F1 Score	Train Accuracy	Test Accuracy
ResNet 50	0.59	0.61	0.64	78%	60%
CNN	0.59	0.61	0.64	71%	64%

Evaluation

Most of the metrics are almost same for both the models, but there is an indication of overfitting in the Train accuracy and Testing accuracy columns as there is larger difference between them in ResNet [18%] as compared to mere [7%] of our CNN model.

Conclusions

Though ResNet50 started with an higher accuracy does not mean that the our CNN model cannot surpass the accuracy of the pre-trained ResNet50 model with enough training epochs and careful hyperparameter tuning.
Higher accuracy may not be the always the appropriate metric of evaluation, here our model has slightly lower accuracy but is more generalized and robust.

Future Scope

Our model is a single label classification model. Further, instead of just training and classifying images, we can develop a multimodal model to classify emotions on multi-channel data like video and audio.