Final Project Report

Madiba Hudson-Quansah

Introduction

The topic of this project is the application of Neural Networks, specifically Convolutional Neural Networks, to the problem of emotion detection in images of faces. Emotion detection is a problem that has been studied in the field of computer vision for decades and has many applications in fields such as health monitoring, human-computer interaction, and market research. This project aims to develop a Convolutional Neural Network based system to detect emotions from pictures using the FER 2013 dataset. The results of this project could potentially be used to enhance automated system's understanding of human emotions, enabling more intuitive interactions and impactful applications.

Methods

Data Collection

The Facial Expression Recognition 2012 (FER 2013) dataset was used for this project, sourced from Kaggle. This dataset contains 35,887, 48x48 grayscale images of faces labelled with one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. This dataset is split into a training and test set of 28,709 and 7,178 images respectively.

Preprocessing

Model Architecture

Baseline Model

- A simple CNN architecture with three convolutional layers followed by a max-pooling layer in the traditional CNN structure, with filters starting at 64 and doubling at each layer. These layers are followed by a feed forward network with two hidden layers and a softmax output layer.
- This baseline model achieved a training accuracy of 25%, a validation accuracy of 25%, and a test accuracy of 24%.

Intermediate Models

In the development of the final model, several intermediate models were tested to improve performance. These models were based on variations of the baseline model with additional layers and alternate layer placements, even including a modified ResNet architecture. However, these models did not improve performance significantly

Final Model

From previous experiments, it was clear that the baseline model and its variations were not sufficiently complex to learn the features necessary to accurately classify the emotions. Taking this into account and my limited compute resources, I decided to used a pre-trained model as the base of my final model and fine-tune it on the FER 2013 dataset. The backbone of the final model is the DenseNet201 architecture, a 401-layer deep CNN with 20.2M parameters and pre-trained on the ImageNet dataset. This architecture was chosen because of its compromise between performance, depth and size.

Building on top of the DenseNet201 backbone, the final model consists of a Global Average Pooling layer instead of Flatten layer to reduce the number of parameters and prevent overfitting, followed by a feed forward network three hidden layers deep with dropout layers in between. The output layer is a softmax layer with seven units, one for each emotion class.

Training

Each model's feed forward network was initialized using the He initialization method to prevent vanishing gradients and used the ReLU activation function with the exception of the output layer which used the softmax activation function. Each model also included batch normalization and dropout layers to prevent overfitting. The final model however differs as it uses the Swish activation function as it has been shown to improve performance over the ReLU activation function especially in deeper networks.

Each model was trained for 30 epochs with a batch size of 64 and incorporated early stopping to prevent overfitting. The final model was also trained for 30 epochs with an additional 10 epochs of fine-tuning with a smaller learning rate.

In training the baseline and most of the intermediate models the AdamW optimizer was used with a starting learning rate of 0.0001 and a learning rate scheduler that reduces the learning rate by a factor of 0.1 when the validation loss plateaus. The final mode was however trained using the Stochastic Gradient Descent (SDG) Optimizer with a starting learning rate of 0.1 and with a similar learning rate scheduler. This change in optimizer was necessary as with the depth of the transfer learning model, the AdamW optimizer was unable to converge to a good solution.

Each model was compiled using the categorical cross-entropy loss function and the accuracy metric this is because the problem is a multi-class classification problem and the categorical labels are one-hot encoded.

Evaluation

Metrics

On evaluation, the final model achieved:

Training Accuracy 78 %

Valdiation Accuracy 65 %

Test Accuracy 65 %

These results show a significant improvement over the baseline model and the intermediate models, with minimal to moderate overfitting. The final model was able to learn the features necessary to classify the emotions in the FER 2013 dataset with a high degree of accuracy. The confusion matrix of the final model shows that the model is able to accurately classify the emotions with the exception of the disgust class which is often misclassified as anger.

Conclusion

This project successfully implemented a CNN-based emotion detection system using the FER 2013 dataset. The DenseNet201-based architecture outperformed the baseline model, achieving robust accuracy and generalization. Insights gained include the importance of data augmentation, fine-tuning pretrained networks, and addressing class imbalance.

Future Directions

This could be used to implement a real time emotion detection system for example the basic one implemented in the notebook.