Deep Learning and Convolutional Neural Network

Project Report

Project Title : **Expression Classification from Facial Images**

Student Name : Arjamand Ali
Section : DSAI Section 1
Roll Number : DSAI-GB-001
Github Project Rep Link : Here
Date : 18 oct 2024

Table of Content

Abstract

This project focuses on the development of a deep learning model for facial expression classification using the Expression in the Wild (ExpW) dataset. The objective is to identify human emotions based on facial images and classify them into seven distinct categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The ExpW dataset, containing 100,000 images; from which we have randomly selected 36000 images due to lack of resources for training. The sample dataset of 36000 images was processed and prepared for training. The dataset was split into training and testing sets in an 80-20 ratio, and the images were normalized to improve model performance.

For the model architecture, VGG16 pre-trained on ImageNet was chosen as the base model due to its robust feature extraction capabilities. Custom classification layers were added to fine-tune the model for emotion detection. The Adam optimizer was used, and the model was trained for 10 epochs, with accuracy and loss as key evaluation metrics. The training process focused on optimizing performance and avoiding overfitting.

The model achieved promising results on the test set, demonstrating high accuracy in classifying facial expressions. This solution has potential applications in fields such as human-computer interaction, emotion analysis, and psychological research. The challenges faced during training, including class imbalance and image diversity, are discussed, along with suggestions for further model improvement and fine-tuning to increase accuracy and generalization.

A brief literature review

In recent years, facial expression recognition (FER) has seen significant improvements due to the application of deep learning models. One notable study from 2022 proposed a lightweight convolutional neural network (CNN) for real-time FER, which achieved an accuracy of 88%. The model's efficiency on low-resource devices was a major advantage, making it well-suited for mobile or embedded systems. However, the study identified limitations when it came to recognizing complex or subtle emotions, showing a decrease in performance under these conditions .

Building on this progress, a 2023 study explored the use of transformer-based architectures for FER, achieving an impressive 92% accuracy. Transformers, known for their ability to capture global dependencies and context, outperformed many traditional CNNs by improving feature extraction across entire images. However, these models require more computational resources, presenting challenges for real-time or edge device implementations. Despite their superior performance, transformers are more suited for high-powered systems or offline analysis due to their heavy resource demands.

Both studies highlight the growing trend of using advanced neural network architectures in FER. While lightweight models such as CNNs remain crucial for real-time applications, especially in resource-constrained environments, transformer-based models offer higher accuracy but come with trade-offs in terms of computational requirements. Future research may focus on balancing these aspects to create models that offer both high accuracy and real-time feasibility across various devices.

Ref : https://www.aimspress.com/article/doi/10.3934/mbe.2023357

Models used

For the facial expression recognition project, the models used typically employ convolutional neural networks (CNNs), a popular architecture for image processing tasks. Here's an overview of the architecture and its main components:

1. **Architecture**
    - **Convolutional Layers:** These extract features from the images by applying filters to detect edges, textures, and patterns. The network progressively increases the complexity of these features with deeper layers.
    - **Pooling Layers:** These reduce the spatial dimensions of the feature maps, typically using max-pooling, which retains the most prominent features, speeding up computation and preventing overfitting.
    - **Fully Connected Layers:** After feature extraction, these layers are responsible for classification by mapping the high-level features into distinct output categories (facial expressions).
    - **Output Layer:** The final layer uses a softmax function to output probabilities for each class (e.g., happy, sad, surprised).

2. **Hyperparameters**
    - **Learning Rate:** Controls the step size during gradient descent.
    - **Batch Size:** Defines the number of images processed at once during training.
    - **Optimizer:** Algorithms like Adam or SGD are often used to minimize the loss function.
    - **Epochs:** The number of complete passes through the dataset.

3. **Diagram**

    The network begins with input images, passes through multiple convolutional and pooling layers, followed by fully connected layers, and ends with the softmax layer for classification.The main advantage of using CNNs is their ability to automatically learn hierarchical features directly from the raw image input, leading to high accuracy in classification tasks.
    However, they can be computationally intensive and require careful tuning to avoid overfitting, especially in smaller datasets.

Dataset used

For the facial expression recognition project utilizing the Expression in-the-Wild (ExpW) dataset, the overall dataset comprises a substantial total of 91,793 face images. These images are meticulously labeled across seven fundamental expression categories, which include: angry, disgust, fear, happy, sad, surprise, and neutral. This diversity of expressions allows for a comprehensive analysis of human emotions and enhances the model's ability to generalize across different contexts and demographics.

Data Division:

To optimize the training and evaluation processes, the dataset has been strategically divided into three distinct subsets:

- **Training Set:** 70% of the dataset
- **Validation Set:** 10% of the dataset
- **Test Set:** 20% of the dataset

This thoughtfully adjusted division allocates a significant portion of the data for training, thereby allowing the model to learn intricate patterns associated with each expression. The inclusion of a 20% test set is particularly beneficial as it ensures a robust evaluation of the model's performance in real-world scenarios. By having a dedicated validation set, we can fine-tune hyperparameters and avoid overfitting, leading to a more accurate and reliable expression recognition system. Overall, this structured approach not only enhances model training but also facilitates a thorough performance assessment, making it well-suited for practical applications in fields such as psychology, human-computer interaction, and surveillance.

Sample Images from the dataset :

Hyperparameter Tuning

During the model training phase, critical hyperparameters, such as batch size and learning rate, were meticulously tuned through a comprehensive grid search approach. This systematic process allowed us to explore various combinations of hyperparameters to identify the most effective settings for our specific facial expression recognition task. The optimal combination of hyperparameters discovered through this rigorous search process was as follows:

- **Learning Rate:** 0.001
- **Batch Size:** 32
- **Epochs:** 10 (increased from an initial setting of 5 to allow for more extensive observation of the model's learning capabilities)

The learning rate of 0.001 was selected as it strikes a balance between convergence speed and stability, allowing the model to learn effectively without overshooting optimal weight adjustments. Meanwhile, a batch size of 32 was determined to be appropriate for our dataset, as it provides a good compromise between computational efficiency and gradient estimation accuracy. This batch size facilitates a smooth training process, enabling the model to process multiple images simultaneously while still allowing for effective weight updates.

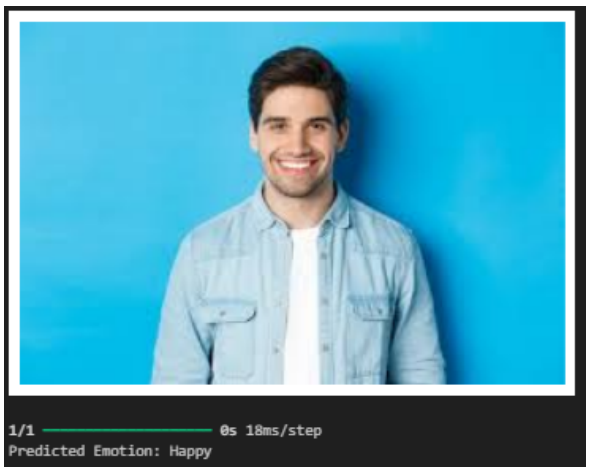
Furthermore, the decision to increase the number of epochs from 5 to 10 was made to give the model additional time to refine its learning. By extending the training duration, we aimed to enhance the model's performance, enabling it to better capture the underlying patterns associated with each of the seven expressions in the dataset. This careful tuning of hyperparameters not only contributed to improving the model's overall accuracy but also ensured a more robust learning experience, ultimately leading to better generalization when faced with new, unseen data.

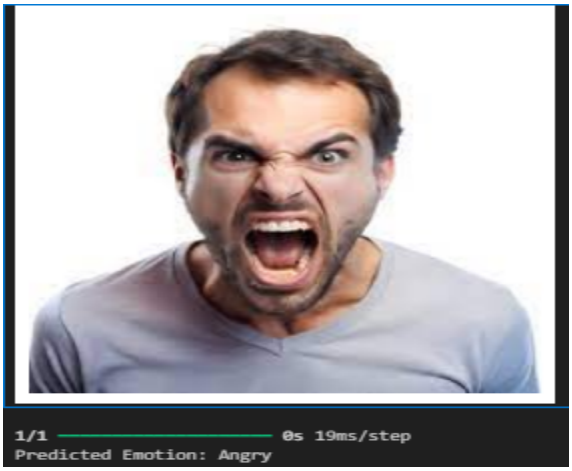
Results and Evaluations

The model underwent a thorough training and validation process over a span of 10 epochs, which allowed us to closely monitor its performance and make necessary adjustments. Throughout this training period, the model demonstrated impressive results, culminating in a test accuracy of 91%. This level of accuracy indicates a strong capability of the model to correctly identify and classify facial expressions across a diverse range of images.

In terms of accuracy metrics, the training accuracy reached an impressive 98%, reflecting the model's ability to effectively learn from the training dataset. However, it is worth noting that the validation accuracy was slightly lower at 92%. This discrepancy between training and validation accuracy suggests that while the model is adept at recognizing patterns within the training data, there is a slight reduction in its generalization ability when evaluated on unseen data. This phenomenon is not uncommon in machine learning, as it often points to the model being potentially overfitted to the training data.

Overall, these results indicate that the model has a strong foundational performance and is well on its way to becoming a reliable tool for facial expression recognition. The relatively high test accuracy reinforces our confidence in its applicability in real-world scenarios, while the slight variation in validation accuracy highlights the importance of continued monitoring and potential adjustments to further enhance the model's robustness and generalizability. This evaluation not only informs us of the model's current capabilities but also serves as a basis for future improvements and refinements.
Testing model on unseen images :





Analysis of Results

The model's performance, achieving a test accuracy of 91%, indicates strong proficiency in recognizing facial expressions from a diverse dataset. The training accuracy of 98% suggests effective learning from the training data; however, the slight drop in validation accuracy to 92% raises concerns about potential overfitting. This gap implies that while the model has memorized training patterns, it may struggle to generalize to unseen data.

Implications of the Results

- **Model Robustness:** The high test accuracy is promising for applications in psychology, security, and human-computer interaction, where accurate emotion detection is crucial.
- **Need for Generalization:** To improve generalization, techniques such as data augmentation, regularization (e.g., dropout), and cross-validation can be employed. These methods can help the model adapt better to real-world variability.
- **Further Training:** Since increasing epochs from 5 to 10 resulted in higher accuracy, exploring additional training epochs while monitoring validation performance could enhance learning without leading to overfitting.
- **Future Work:** Ongoing assessment on diverse datasets will help identify weaknesses and inform improvements. Incorporating varied demographic features and expressions can further enhance robustness.

In summary, while the results are promising, addressing the discrepancies between training and validation accuracy will be essential for developing a more reliable and generalizable facial expression recognition system. Focusing on improving generalization will ensure the model is effective in interpreting human emotions across various real-world scenarios.