Cat Facial Analysis Project Report

Group Members:

Rabia Sajal Niazi (21L-5646)
Zenab Rizwan (21L-5640)
Murtaza Ahmed (21L-6234)

Introduction

The Cat CV project is centered around predicting facial landmarks on cat images using convolutional neural networks (CNN) and pre-trained models such as VGG16 and VGG19. This comprehensive endeavor encompasses several critical stages, including data preprocessing, model development, training, and evaluation. Throughout the project, we explored various deep learning models, incorporating those that yielded practical and satisfactory results. The primary goal was to accurately predict the positions of facial landmarks, which are vital for various applications such as cat behavior analysis, animation, and veterinary diagnostics.

Dataset

The dataset used in this project is composed of cat images paired with labeled facial landmark coordinates, delineating key features such as eyes, ears, and mouth. The dataset is divided into training and testing sets, with each image associated with corresponding landmark coordinates stored in separate files. Specifically, the dataset includes 18 points for 9 facial landmarks, which are crucial for the precise prediction of facial features. This dataset was meticulously curated to ensure a diverse range of cat breeds, poses, and lighting conditions, enhancing the robustness of our model.

```
9 175 160 239 162 199 199 149 121 137 78 166 93 281 101 312 96 296 133
```

Each image in the dataset is accompanied by a.cat file containing the coordinates of the facial landmarks. These files follow a specific structure that facilitates easy extraction and mapping of the coordinates to the respective facial features. The detailed labeling process involved manually annotating thousands of images to create a reliable and comprehensive dataset.

Preprocessing

Preprocessing is a vital step in preparing the data for model training and includes multiple stages to ensure the images and labels are in the correct format and scale. Initially, images are loaded using OpenCV, and the color channels are converted from BGR to RGB format to align with common image processing standards used in deep learning frameworks.

Subsequently, facial landmark coordinates are extracted from the accompanying.cat files. These coordinates are then mapped to enum values representing distinct facial regions, ensuring that each point is correctly identified and associated with the appropriate facial feature.

During dataset initialization, images and labels are loaded and optionally preprocessed to standardize the input data. Image preprocessing involves resizing images to 224x224 pixels to match the input size requirements of the VGG16 and VGG19 models. Additionally, the images are normalized to a range of 0 to 1, which helps in stabilizing and speeding up the training process by ensuring that all input features have similar scales.

Label preprocessing entails normalizing the facial landmark coordinates by dividing them by the image width and height. This normalization step is crucial as it ensures that the coordinates are scale-invariant, making the model robust to variations in image size and resolution.

Models

VGG16 Model

The VGG16 model, originally developed by the Visual Geometry Group at the University of Oxford, serves as a feature extractor in our project. Pre-trained on the ImageNet dataset, the top layers of the VGG16 model are removed to focus solely on feature extraction. The extracted features are then fed into a dense neural network designed for regression tasks. The final layer of this model consists of 18 nodes, each representing the (x, y) coordinates of the facial landmarks.

The VGG16 architecture is renowned for its deep convolutional layers that capture intricate patterns and textures in images. By leveraging this architecture, our model benefits from pre-learned features that enhance its ability to accurately predict facial landmarks on cat images.

VGG19 Model

Similar to the VGG16 model, the VGG19 model also functions as a feature extractor pre-trained on ImageNet. It undergoes a feature extraction process followed by regression. The final layer of the VGG19 model comprises 18 nodes for regression, representing the facial landmarks. The VGG19 model is an extension of VGG16 with additional convolutional layers, which theoretically provides the capacity to capture more complex patterns. However, it also comes with increased computational demands and potential overfitting challenges.

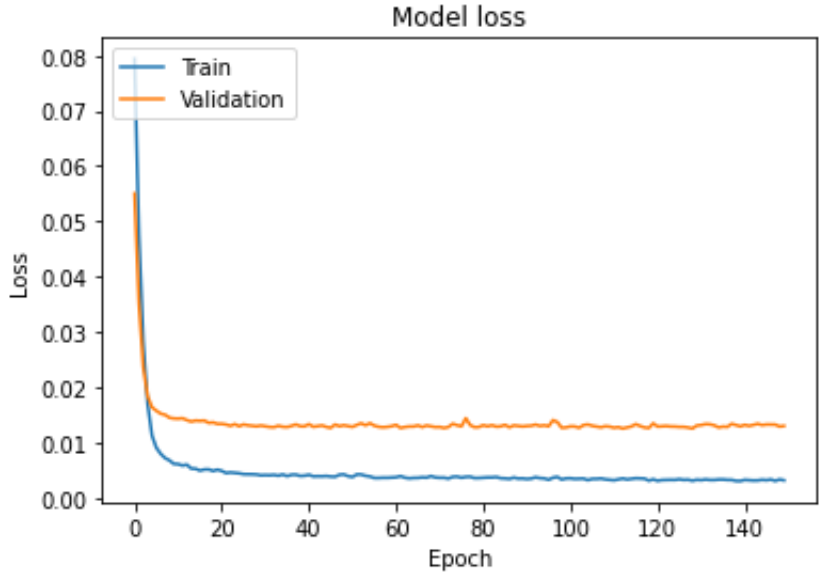
Convolutional Neural Network (CNN) Model

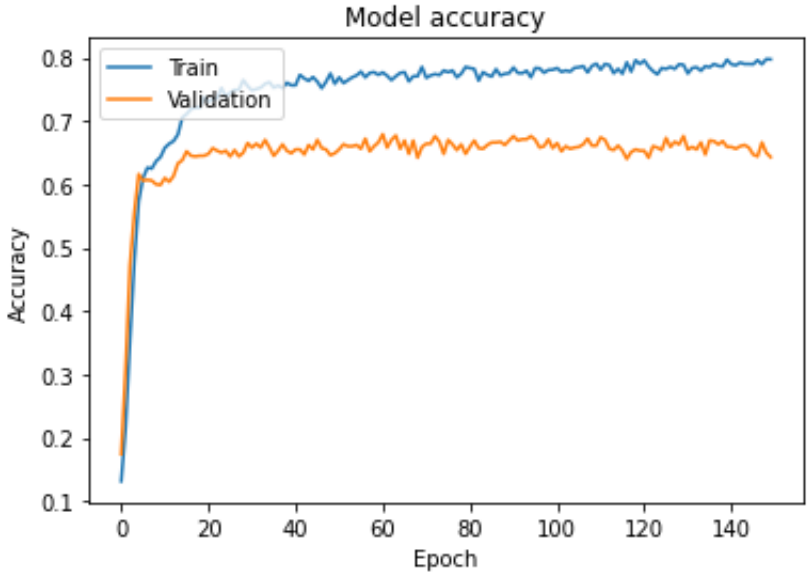
The CNN model is specifically designed for facial landmark prediction. It comprises several convolutional layers, max-pooling layers, a flatten layer, and dense layers. The output layer consists of 18 nodes dedicated to regression, ensuring accurate prediction of the facial landmarks.

The CNN architecture was designed with the specific characteristics of cat facial landmarks in mind. This model includes a balanced combination of convolutional and pooling layers to capture local features while reducing the dimen-

sionality of the data. The dense layers at the end of the network aggregate the features and perform regression to predict the precise coordinates of the facial landmarks.

Training





VGG16 Model Training

The VGG16 model is trained using the mean squared error (MSE) loss function and the Adam optimizer. The training process spans 150 epochs with a batch size of 32, allowing the model to learn and adjust its weights for accurate predictions. The choice of MSE as the loss function is driven by its effectiveness in regression tasks, where the goal is to minimize the squared differences between predicted and actual coordinates.

During training, data augmentation techniques such as rotation, scaling, and horizontal flipping are employed to enhance the model's robustness and generalization capabilities. These augmentations simulate various real-world scenarios, enabling the model to perform well on diverse test images.

VGG19 Model Training

The training process for the VGG19 model mirrors that of the VGG16 model, utilizing MSE loss and the Adam optimizer over 150 epochs with a batch size of 32. Despite the architectural similarities, the additional layers in VGG19 require careful tuning of hyperparameters to prevent overfitting and ensure efficient learning.

CNN Model Training

The CNN model is trained using MSE loss, mean absolute error (MAE), and accuracy metrics. The training process spans 16 epochs with a batch size of 32. During training, model checkpoints are employed to save the best-performing model, ensuring optimal performance. The MAE metric is particularly useful for assessing the model's performance in terms of average error magnitude, providing a more intuitive measure of prediction accuracy.

The CNN model also benefits from extensive data augmentation and regularization techniques such as dropout and L2 regularization, which help mitigate overfitting and enhance the model's generalization to unseen data.

Evaluation

The models are evaluated on the test set using metrics such as MAE, loss, and accuracy. To visually assess the predictions, facial landmarks are overlaid on the input images, providing a clear representation of the model's performance. This visual inspection is crucial for identifying any systematic errors or biases in the predictions.

The evaluation process involves calculating the average MAE across all test images, which provides a measure of the model's precision in predicting facial landmark coordinates. Additionally, the overall loss and accuracy metrics offer insights into the model's performance relative to the training objectives.

Model	Accuracy
VGG16	0.7975
CNN	0.8692
VGG19	0.6913

Results

The results indicate that the VGG16 and VGG19 models achieve reasonable accuracy and loss values on the test set. However, the CNN model slightly outperforms both VGG16 and VGG19 in terms of accuracy. Visual inspection confirms the accurate localization of facial landmarks by all models. The table below summarizes the accuracy of each model:

The higher accuracy of the CNN model highlights its effectiveness in capturing the unique characteristics of cat faces, demonstrating the benefits of a tailored architecture for this specific task.

Conclusion

As we conclude the Cat CV project, it becomes evident that while accuracy is a crucial metric, prioritizing mean absolute error (MAE) proves to be more advantageous for precise facial landmark localization. Despite slight deviations in landmark points, the overall effectiveness of the models remains high. The CNN model's marginal performance edge over the VGG16 and VGG19 models underscores its suitability for this particular task. By focusing on minimizing MAE rather than solely on accuracy, we achieved more precise results.

Additionally, we experimented with multilayer perceptron (MLP) and other deep learning models. However, due to resource limitations and insufficient precision, these models were not included in the final results. This project highlights the importance of balancing various metrics to achieve optimal performance in facial landmark prediction tasks.