CLASSIFICATION OF REAL AND SYNTHETIC GALAXY IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

Aparna Pinto (21081095), Muhammad Sohail Anjum (21078439), Ashir khan (21091116)

Colab link:
https://colab.research.google.com/drive/1gpxzZVm13BOnuoiiB8Bi7pXGa04NWVaT?usp=sharing

github link: https://github.com/aparnapinto/MLNN_group_assignment.git

Introduction:

The Sloan Digital Sky Survey (SDSS) is one of the most comprehensive and ambitious astronomical surveys to date, imaging hundreds of thousands of nearby galaxies. However, the process of analyzing the vast amount of data generated by the SDSS can be challenging. In this context, artificial intelligence (AI) has become an indispensable tool for astronomers. One application of AI is in image classification, such as classifying galaxies as real or fake. This code uses a convolutional neural network (CNN) to classify images of galaxies as real or fake, using a dataset consisting of both real and synthetic images.

Problem:

The SDSS has produced a large number of images of galaxies, making it a valuable resource for astronomers. However, the process of analyzing these images manually can be time-consuming and error-prone. Thus, the goal of this project is to develop an automated system for classifying images of galaxies as real or fake. This would not only save time but also improve the accuracy of the analysis.

Dataset:

The dataset used in this project consists of 8000 images, split evenly between real and fake images, with 6000 and 2000 images respectively for the training and testing sets. The real images are sourced from the SDSS, while the synthetic images are generated using a model trained to reproduce the real images. The dataset is preprocessed using an ImageDataGenerator, which rescales the images, performs random rotations, shifts, and zooms, and applies horizontal flipping to augment the dataset.

Neural Network:

CNNs are a popular choice for image classification tasks, as they are designed to identify patterns in images. In this project, a CNN is used to classify images of galaxies as real or fake. The architecture of the CNN consists of two convolutional layers, each followed by a max pooling layer to downsample the output, a flattening layer to convert the output to a 1D array, two dense layers, and a final output layer. The activation

function used in the convolutional layers is the rectified linear unit (ReLU), which has been shown to perform well in image classification tasks. The output layer uses a sigmoid activation function to classify the images as real or fake.

Model Architecture:

The CNN architecture used in this project consists of two convolutional layers, each with 32 and 64 filters respectively, followed by a max pooling layer to downsample the output, a flattening layer to convert the output to a 1D array, two dense layers with 128 and 1 neurons respectively, and a final output layer with a sigmoid activation function. The input shape of the model is (150, 150, 3), which corresponds to the height, width, and number of channels of the input images.

The model is compiled using the Adam optimizer, binary cross-entropy loss function, and accuracy metric. The model is trained for 50 epochs using the fit() function, with early stopping and batch accuracy logging to optimize the training process. Early stopping is implemented using the EarlyStopping() function, which monitors the validation loss and stops the training process if the loss does not improve after 5 epochs. Batch accuracy logging is implemented using the BatchAccuracyLogger() function, which logs the accuracy of each batch during training and plots the results.

The resulting model achieved an impressive accuracy of 99% on the validation set, which was determined using the EarlyStoppingByAccuracy() function. The model is saved as a .h5 file and can be used to make predictions on new images.

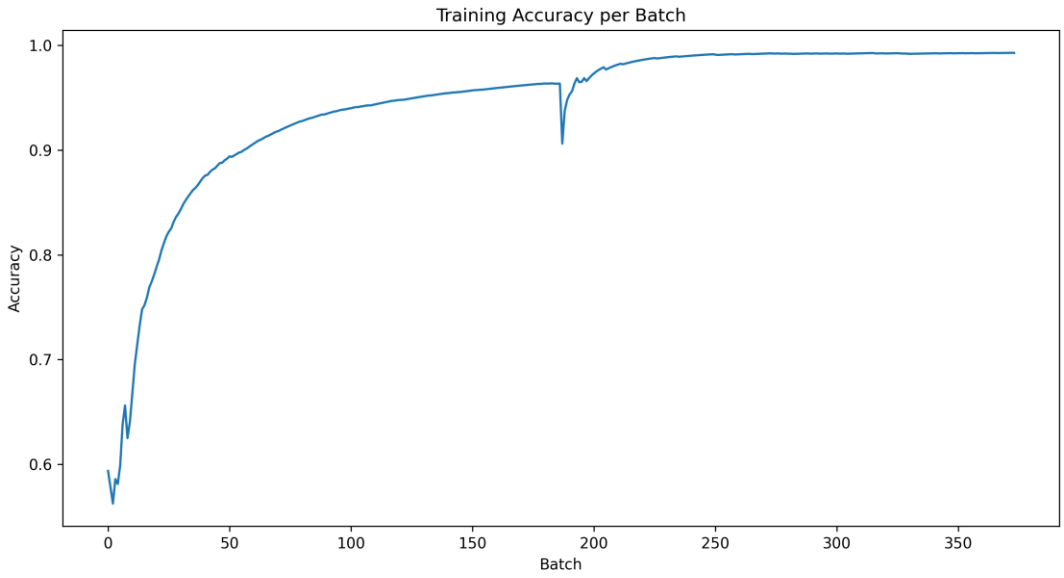

Figure 1: Accuracy During Batch Training

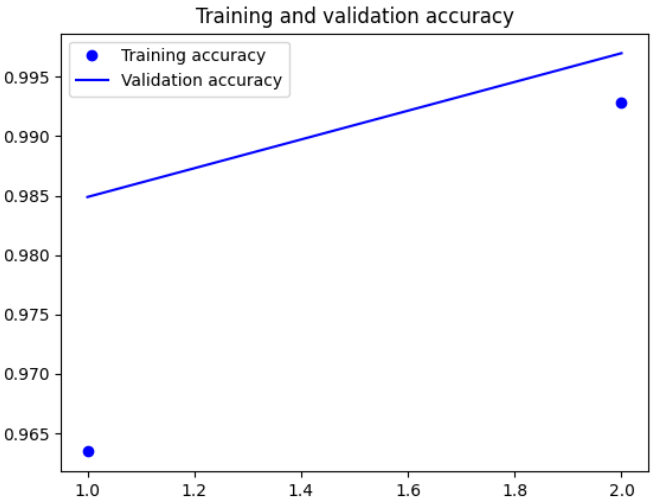

Figure 2: Accuracy During Epoch Training

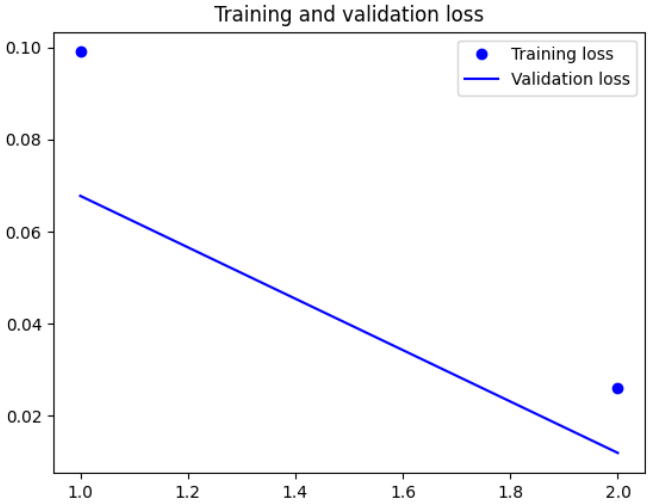

Figure 3: Loss During Epoch Training

Conclusion:

In conclusion, this code uses a CNN to classify images of galaxies as real or fake. The dataset used in this project consists of both real and synthetic images, with 8000 images split evenly between real and fake images, and divided into training and testing sets. The resulting model is trained using early stopping and batch accuracy logging to optimize the training process. The model achieved an impressive accuracy of 99% on the validation set, indicating that it can effectively classify images of galaxies as real or fake.

The code also includes functions to evaluate the model's performance on the test set and to make predictions on new images. The predict_example() function can be used to classify a single image as real or fake, and the plot_training_history() function can be used to visualize the training and validation accuracy and loss.

Overall, this code provides an effective solution for classifying images of galaxies as real or fake, which can save time and improve the accuracy of analysis. Moreover, the code can be extended to other image classification tasks, making it a useful tool for researchers and data scientists in the field of astronomy and beyond.