INITIAL DATA EXPLORATION REPORT

By Blessing Ilesanmi

3Signet Data Science Internship Week 7

8th November, 2024

DATA OVERVIEW AND STRUCTURE

The dataset loaded is the CIFAR-100, which contains images across 100 different classes, each in a 32x32 RGB format, covering a diverse range of object categories, including animals, plants, and various objects (such as apple, bus, rose, and train). The training data comprises 50,000 images, each with dimensions of 32x32 pixels and 3 color channels (RGB). Correspondingly, the labels for the training data have a shape of (50000, 1), with each image having a single label represented as a one-dimensional array. The test data set includes 10,000 images, structured identically to the training data with a shape of (10000, 32, 32, 3) for images and (10000, 1) for labels. Each image is of 32x32 pixels, in RGB color, and has a data type of uint8, representing pixel values from 0 to 255. Labels are of data type int32, which is suitable for encoding class labels.

- **Training Images Shape**: (50,000, 32, 32, 3)
- **Training Labels Shape**: (50,000, 1)
- **Test Images Shape**: (10,000, 32, 32, 3)
- **Test Labels Shape**: (10,000, 1)

DATA QUALITY AND RANGE ANALYSIS

The pixel values in the CIFAR-100 dataset range from 0 (black) to 255 (white), confirming that the images are in an 8-bit format. The mean pixel value is approximately 121.94, indicating that the dataset's brightness centers around a mid-range value. The standard deviation of pixel values is around 68.39, reflecting a wide range of pixel intensities across the images. A check for missing values revealed none in either the training images or labels, ensuring that the dataset is complete and ready for modeling without any need for imputation.

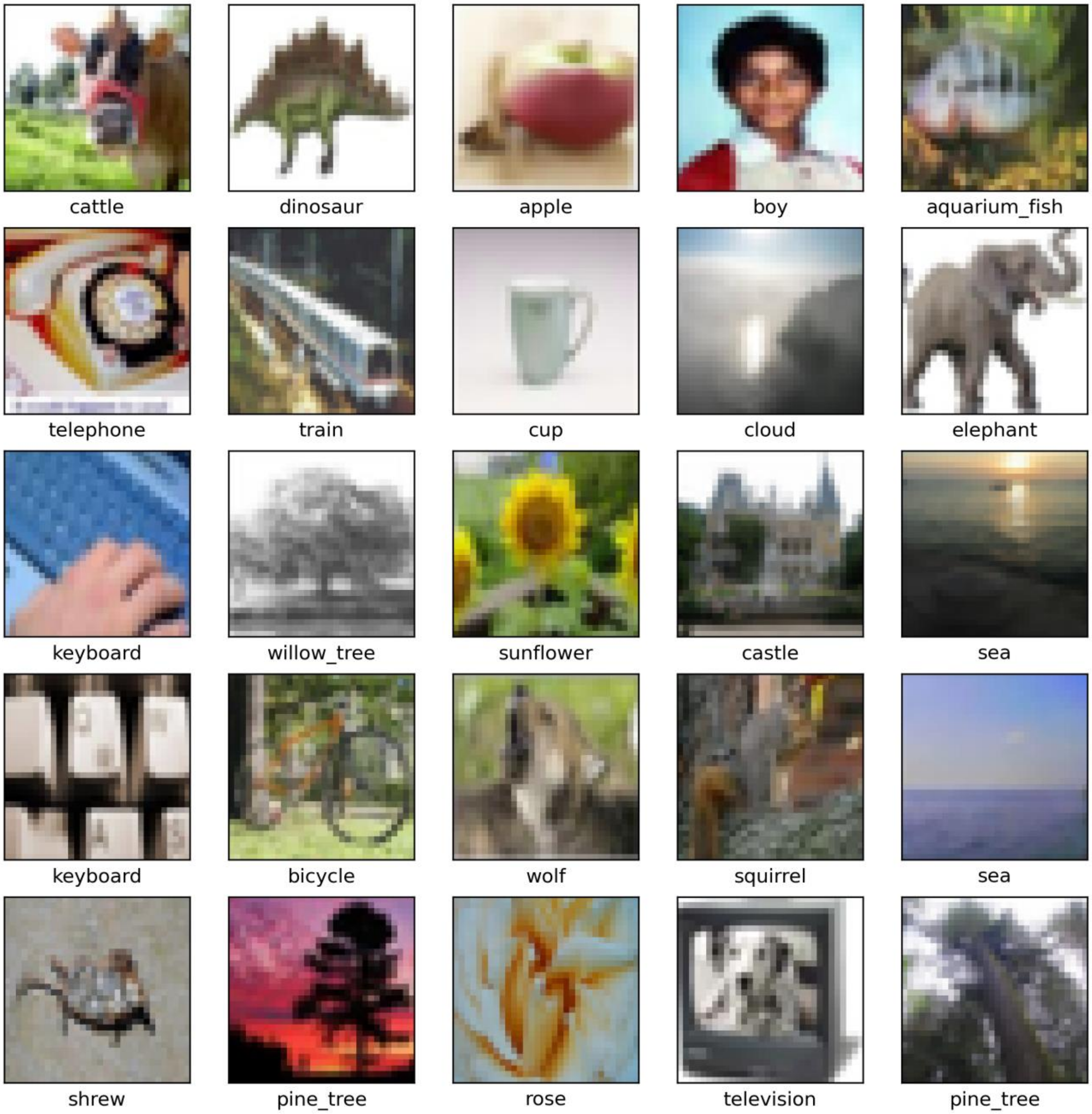
- **Minimum Pixel Value**: 0
- **Maximum Pixel Value**: 255
- **Mean Pixel Value**: 121.94
- **Standard Deviation of Pixel Values**: 68.39

VISUAL INSPECTION OF SAMPLE IMAGES

A subset of 25 random training images was displayed with their corresponding labels, providing a snapshot of the dataset. The images span a diverse range of classes, giving an impression of the variety and quality of data.

This visualization provided:

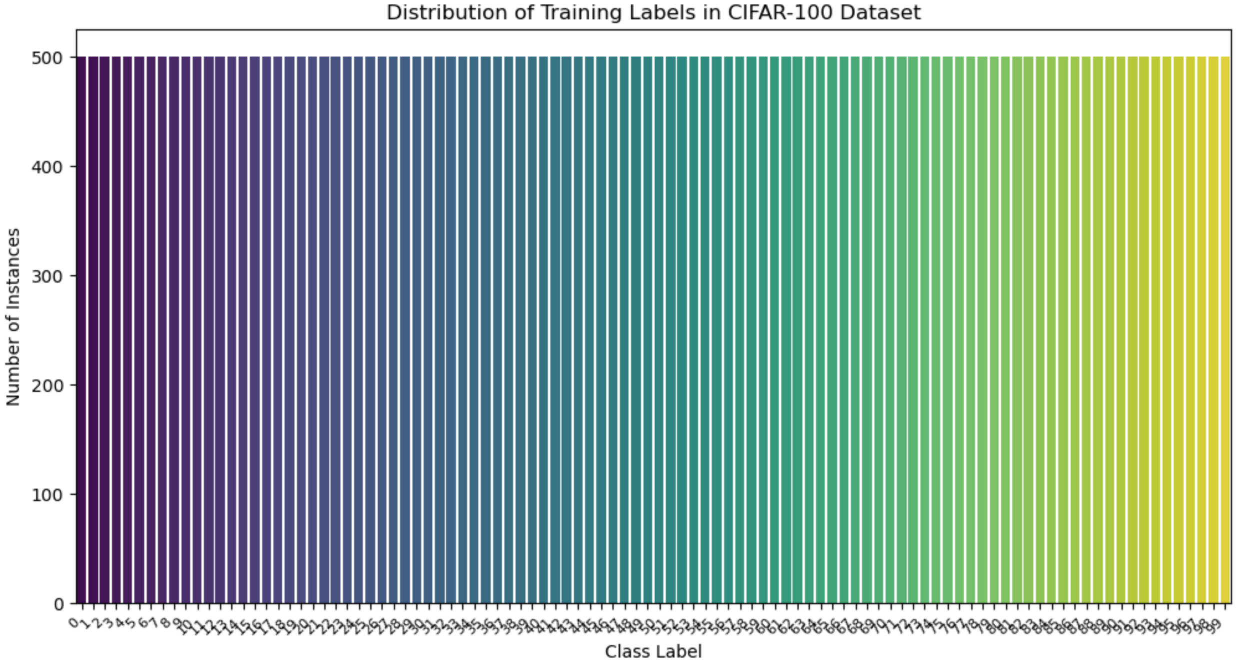
- A quick visual confirmation of the diversity in the CIFAR-100 dataset.
- Examples of various objects and entities that the model will learn to classify.

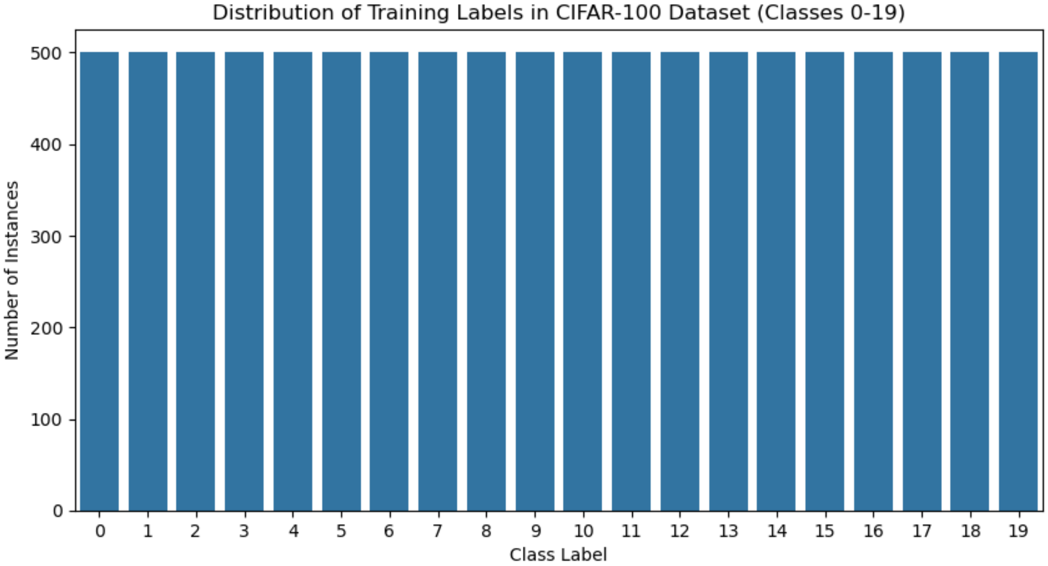


Observation: The images are relatively small in resolution (32x32), which can be challenging for models to recognize complex features but is manageable with convolutional neural networks (CNNs) designed for image data.

LABEL DISTRIBUTION IN TRAINING SET

The distribution of labels in the training set was plotted, showing the count of instances per class. Each of the 100 classes has approximately the same number of instances in the training set, confirming that the CIFAR-100 dataset is balanced in terms of label distribution. This balance is advantageous for model training, as it helps to reduce bias toward certain classes and can improve the model's generalization across all classes.





Due to the clumsiness of visualizing all 100 classes in a single chart, the data was divided into groups of 20 (The 2nd chart is showing the first group).

CONCLUSION

This initial data exploration of the CIFAR-100 dataset reveals that it is a well-organized, balanced, and complete dataset, ideal for training and evaluating image classification models. The dataset's diverse classes, consistent distribution across labels, and clean data structure make it suitable for deep learning tasks without requiring extensive preprocessing. Future steps in model development can focus directly on building and optimizing the model, as the dataset's characteristics suggest minimal risk of bias or skewed performance across classes. This analysis confirms that the CIFAR-100 dataset is ready for further processing and experimentation in machine learning applications.