DATA EXPLORATION AND PREPROCESSING REPORT

By Blessing Ilesanmi

15th November, 2024

INTRODUCTION

This report provides an in-depth exploration and preprocessing of an image dataset intended for training a machine learning model. The dataset contains images from 100 distinct classes, and the goal is to transform and prepare the data to improve the efficiency, accuracy, and generalization capability of the model. The preprocessing steps involve exploring the distribution of labels, normalizing the image data, applying data augmentation to artificially increase the dataset size, and visualizing a subset of the data for quality inspection.

LABEL DISTRIBUTION

The dataset consists of 100 unique classes, and upon examining the distribution of labels, I observed that each class has exactly 500 images. This balanced distribution is a critical factor for the success of training a machine learning model, as it ensures that the model is not biased toward any particular class. With a balanced dataset, the model is less likely to overfit on classes that are over-represented and will be encouraged to learn to identify characteristics that are generalizable across all classes.

By counting the number of images per class, I confirmed that the dataset is evenly distributed, which is essential for training deep learning models. A uniform class distribution mitigates the risk of model bias and provides the model with enough data from each class to learn meaningful features.

DATA NORMALIZATION

The original image data consisted of pixel values ranging from 0 to 255. However, for most machine learning models, especially deep learning models, it is essential to scale the image data to a normalized range. Normalization helps the model by transforming the pixel values into a range that is more suitable for training.

In this case, I performed a simple rescaling operation where the pixel values were converted from a range of [0, 255] to [0, 1] by dividing the pixel values by 255. This rescaling ensures that the values are small and within a standard range, which can significantly speed up the convergence of gradient-based optimization algorithms used in training. Normalization also ensures that the model doesn't treat features with larger numerical ranges (like pixel values in a 0-255 range) as more important than features with smaller ranges.

The normalization step reduces the model's reliance on the scale of the input features, helping it to generalize better to unseen data.

DATA AUGMENTATION

Data augmentation is a key technique used to artificially increase the size of the dataset and introduce variety into the images. This is particularly important when the original dataset is relatively small or lacks diversity in terms of pose, lighting, or other conditions that might be

encountered in real-world use cases. Augmentation helps the model become more robust by allowing it to learn from a wider range of slightly modified versions of the same images.

In this case, several augmentation techniques were applied to each image in the dataset:

1. **Rotation**: Random rotations of the images by up to 15 degrees were applied. This is useful to make the model invariant to slight rotations of objects in the images.
2. **Shifting**: Horizontal and vertical shifts (up to 10% of the image width or height) were applied to simulate slight movements of the objects within the image, such as objects being slightly off-center.
3. **Flipping**: Horizontal flipping was applied to simulate different orientations of the object (e.g., flipping a left-facing object to appear right-facing).

By applying these transformations, I increased the diversity of the training set, allowing the model to learn invariant features, which helps improve its performance when encountering new, unseen data. For each class, 500 augmented images were generated, increasing the overall dataset size significantly.

DIRECTORY STRUCTURE FOR AUGMENTED DATA

A crucial part of the augmentation process was saving the augmented images in an organized directory structure. For each class, a dedicated subdirectory was created, and all augmented images belonging to that class were stored in the respective folder. This approach ensures that the labels of the augmented images remain consistent with the original dataset.

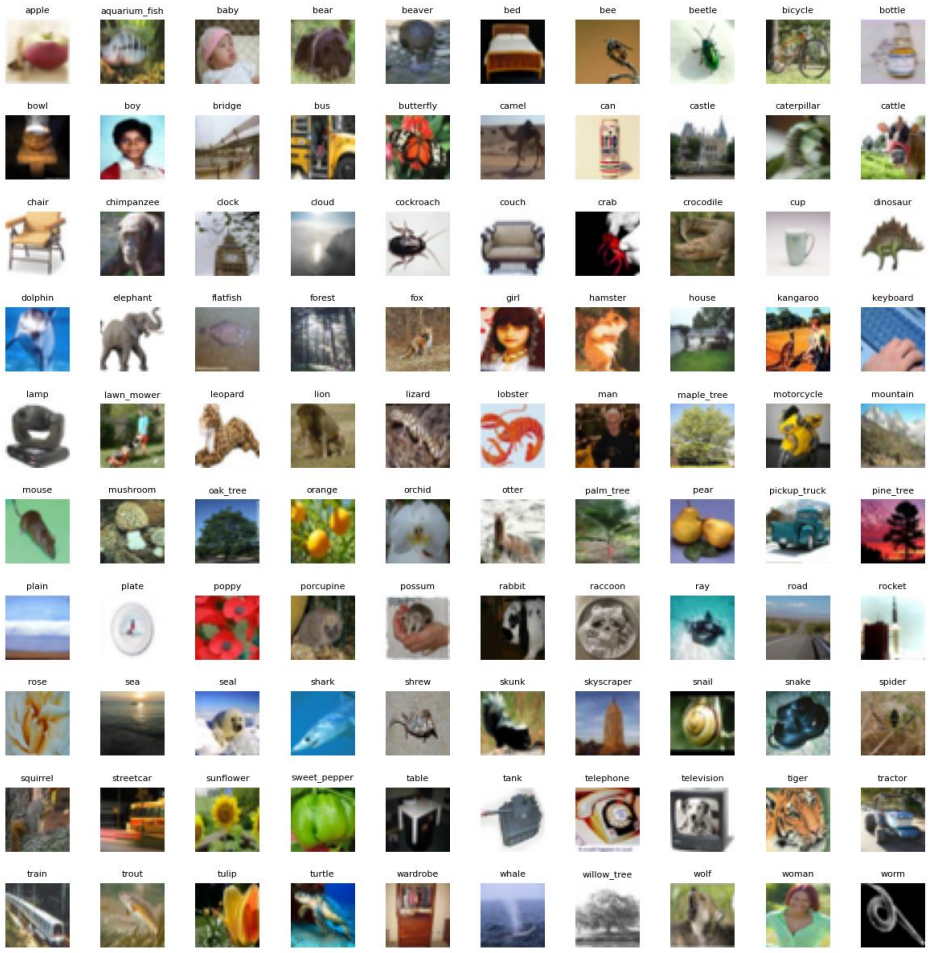
The images were saved using their class name as the prefix for the filename, ensuring that the augmented images were properly labeled and could be easily identified later. This organizational structure is important for maintaining label integrity during training and for any future data analysis or model evaluation.

SAMPLE VISUALIZATION OF THE DATASET

To inspect the dataset visually, a random subset of 100 images was selected and displayed in a grid. These images were carefully chosen from 10 different classes, with 10 images shown per class. This visual inspection serves several purposes:

- **Label Verification**: Ensures that the images are correctly labeled according to their class. This step is essential to catch any labeling errors that may have occurred during the data collection process.
- **Quality Check**: Allows for a visual check of the diversity of the dataset. This helps confirm that the dataset contains a wide variety of images from each class, which is critical for training a robust model.
- **Content Inspection**: The visualized images provide insight into the characteristics of the images in the dataset, such as their complexity, resolution, and content diversity. This helps identify potential challenges the model might face during training (e.g., images with low contrast or poor resolution).

By plotting these 100 images, I ensured that the data was of high quality, correctly labeled, and sufficiently diverse to represent the real-world variations the model will encounter.



CONCLUSION

The data exploration and preprocessing steps have successfully prepared the dataset for training a machine learning model. By normalizing the images, augmenting the dataset, and ensuring that the class distribution remains balanced, we have addressed several common challenges in machine learning. The data is now well-structured, diverse, and ready for the next phase of the model development pipeline.

These preprocessing steps have set a strong foundation for training a model that can generalize well to unseen data. The augmented data will help mitigate overfitting, and the normalization will

ensure that the model converges efficiently. With these preparations in place, we can move forward to training the model and evaluating its performance on new, unseen data.