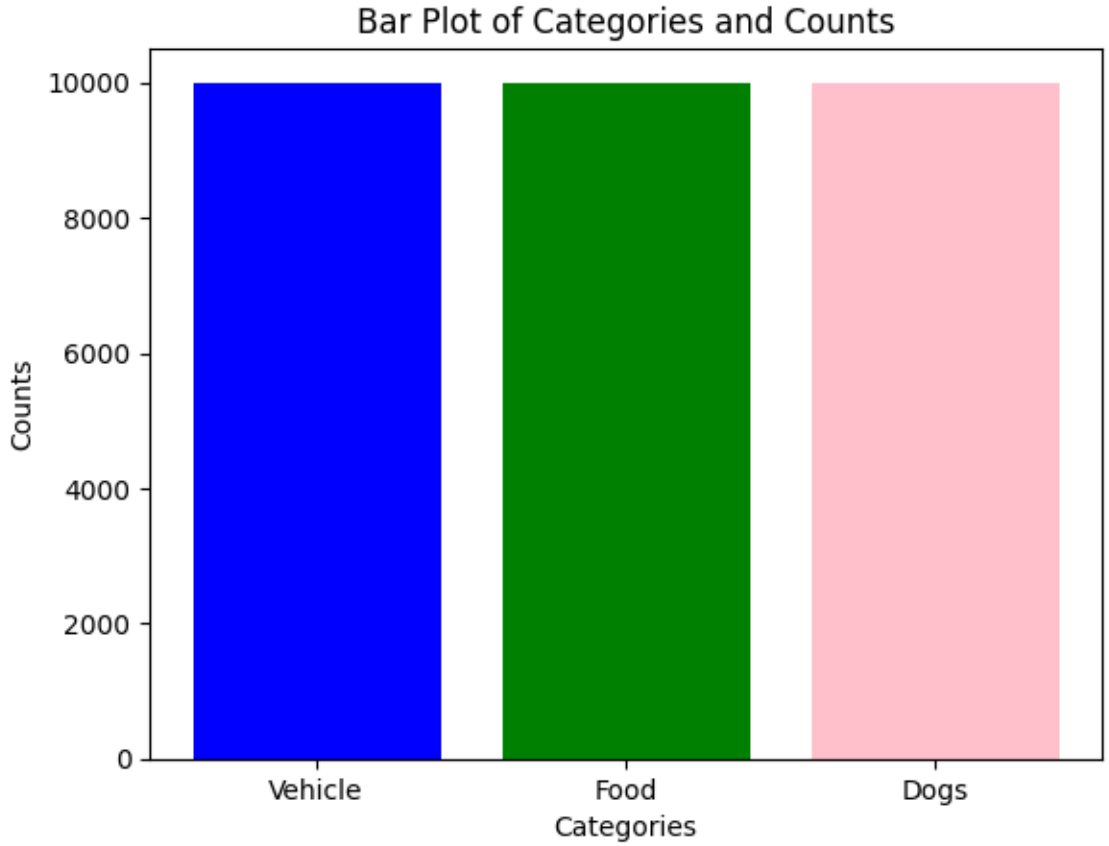
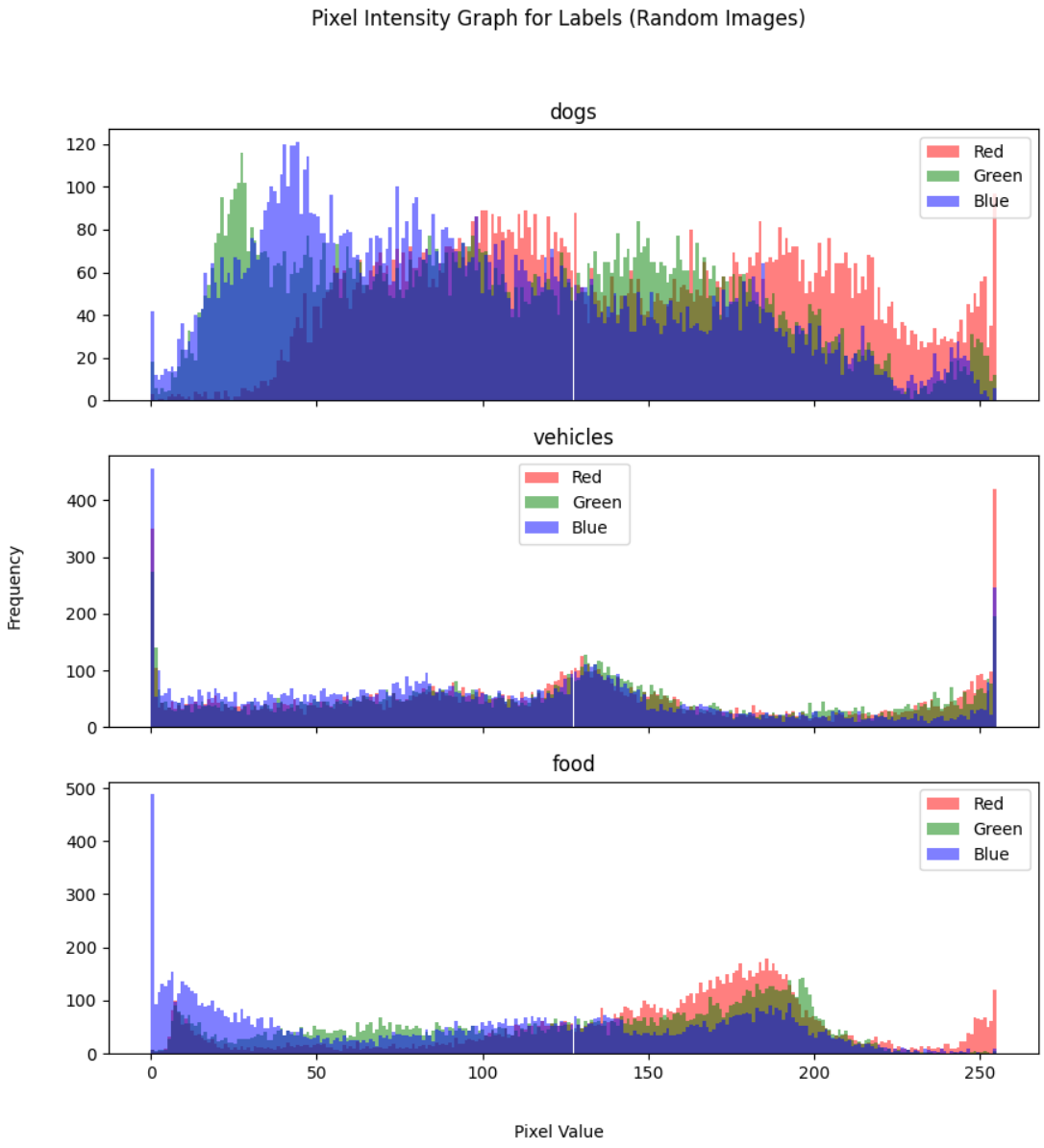
Part III: Implementing & Improving AlexNet

1) The dataset at hand is comprised of three distinct categories: food, vehicles, and dogs. Each classification is composed of 10,000 images that are 64 by 64 pixels in size and in the RGB format. The dataset is also equipped with key statistical measures such as minimum and maximum pixel values, mean and standard deviation of pixel values, as well as the distribution of classes. Notably, the minimum pixel value recorded is zero, the maximum is 255, the mean pixel value hovers around 112, and the standard deviation is approximately 70.



The above Figure is a bar plot of distribution of the classes in the training set, which shows the count of the number of samples on the y-axis and the class labels on the x-axis. It is clear from the figure that each label has equal number of instances.

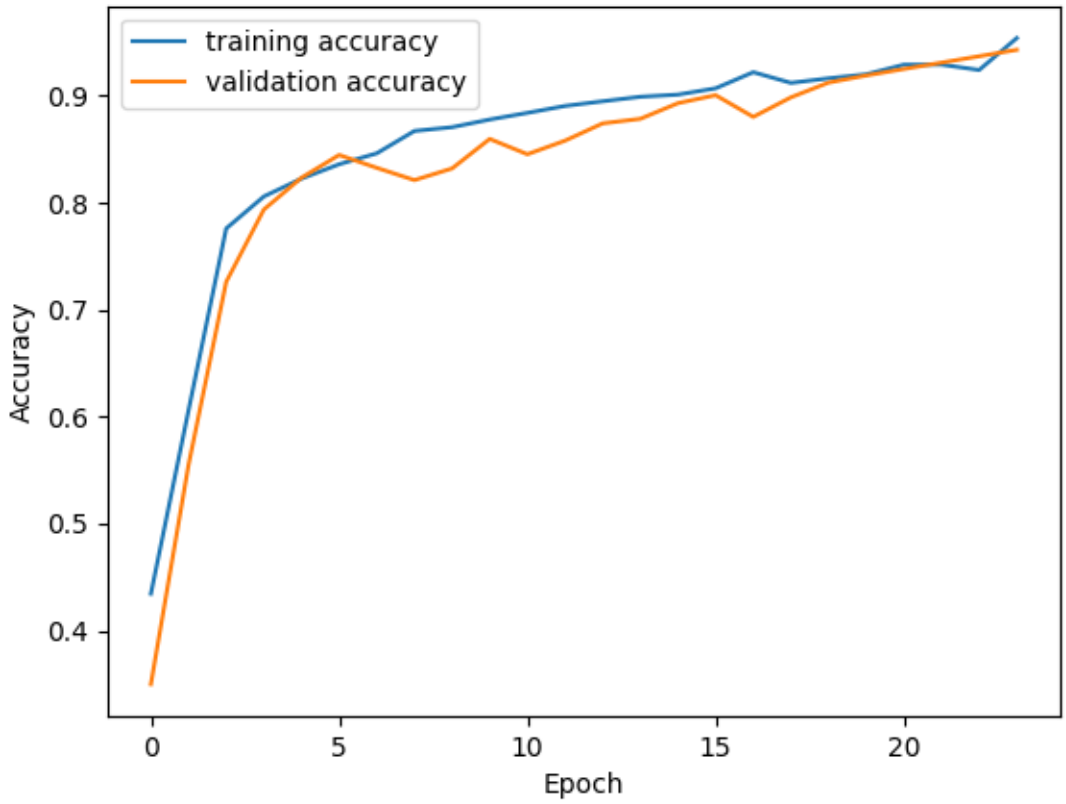
The above figure is a 1X9 graph plotting random images along with their labels from the dataset to get a sense of what the images look like.



The above figure is of a pixel intensity graph which shows the frequency of occurrence of each pixel value on the x-axis, and the number of pixels with that value on the y-axis. We have pixel intensity value for each of the three labels ‘dogs’, ‘vehicles’, and ‘food’.

2) For the task of image classification, we employed the renowned convolutional neural network architecture, 'AlexNet'. This model is composed of 5 convolutional layers, which are followed by

3 fully connected layers. Additionally, it is equipped with max-pooling layers, dropout regularization, and ReLU activation functions. Notably, the final layer of the model comprises 3 units, which align with the 3 distinct labels that are required for classification.



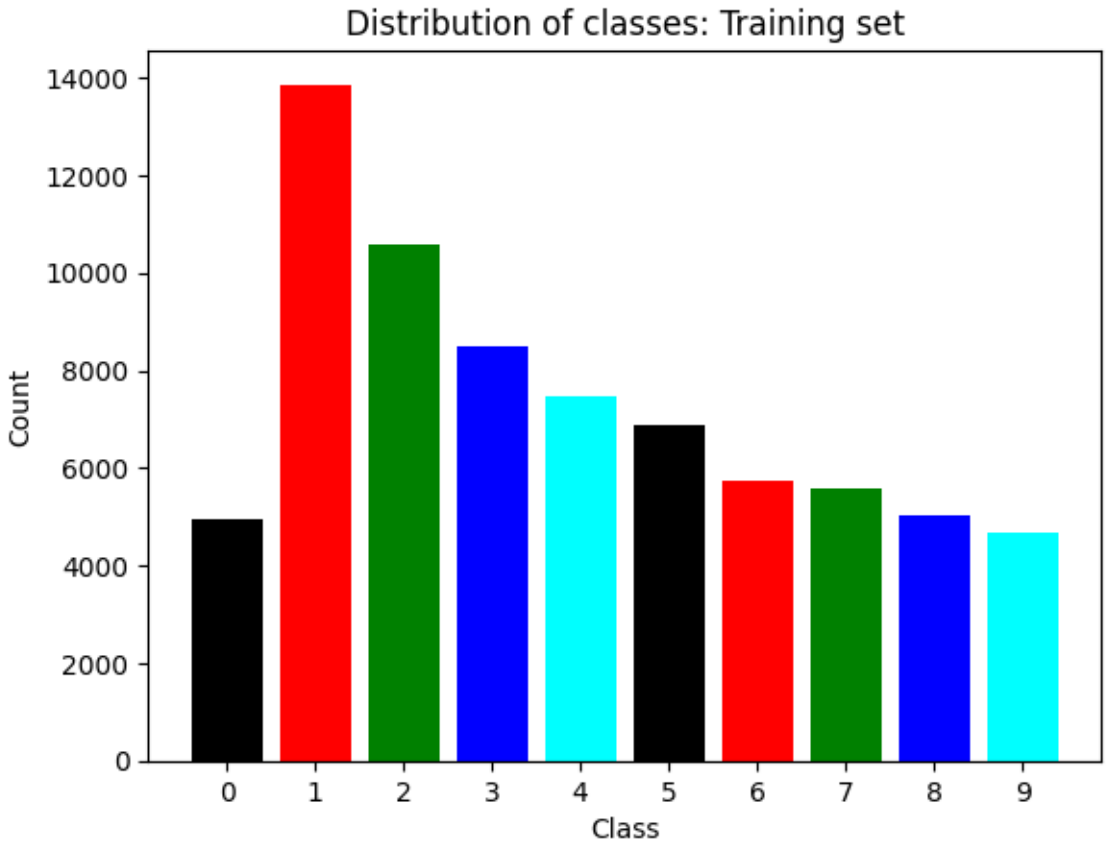
The figure presented above displays a comparison of the test and training accuracy of the AlexNet model on the given dataset. The results indicate that the model is demonstrating excellent performance and is avoiding overfitting the data, which has enabled us to attain an accuracy of over 94%.

3) We implemented several modifications to enhance the performance of the AlexNet model. Firstly, we introduced additional dropout and batch normalization layers to prevent overfitting. We also decreased the number of filters in certain convolutional layers, which resulted in similar, if not improved, performance compared to when more filters were used. Additionally, we adjusted the learning rate during training to expedite the training process on the training set.

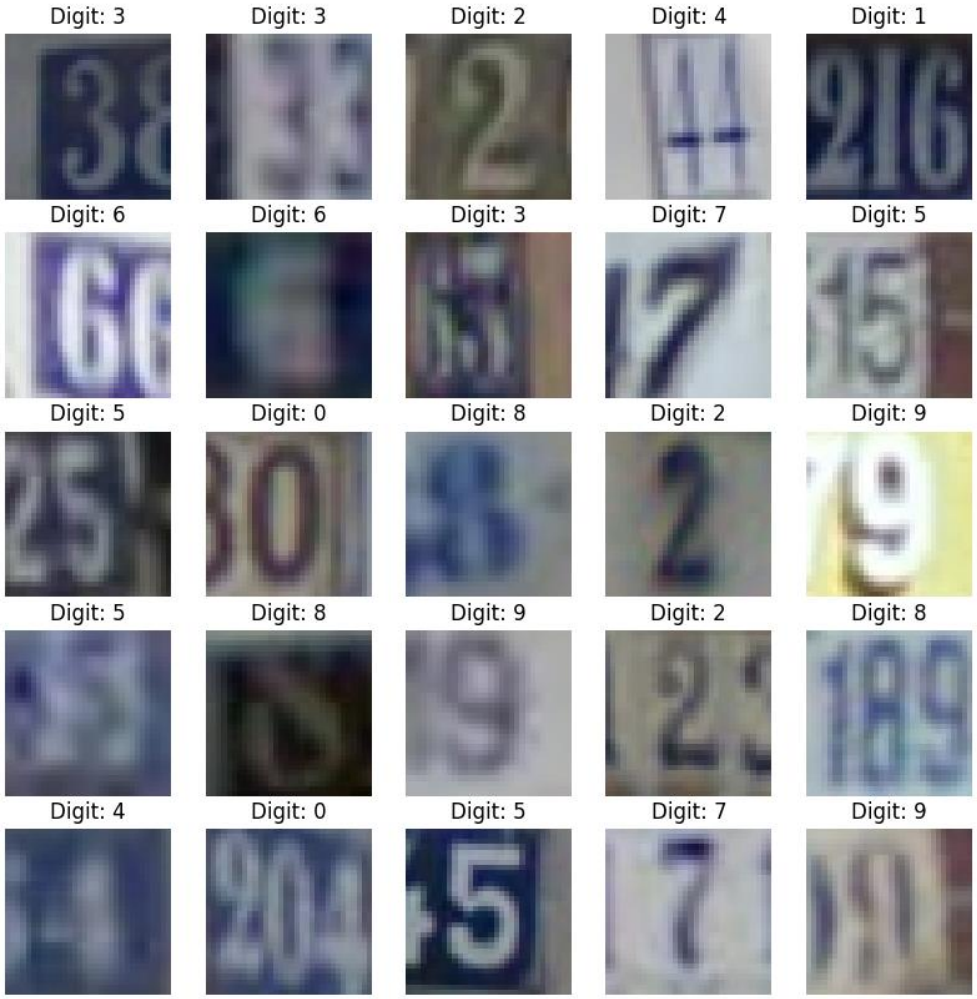
Part IV: Optimizing CNN + Data Argumentation

1) The Street View House Numbers (SVHN) dataset is a real-world image dataset used for developing machine learning and computer vision algorithms and models. It consists of more than 600,000 digit images of house numbers taken from Google Street View. The data in the dataset consists of 32X32 images which are in RGB format (3 Channels). Each image in the dataset contains a single digit from 0 to 9, and the task is to classify each digit correctly. The dataset is split into three parts: a training set with 73257 images, a validation set with 26032 images, and a test set with 26032 images. Corresponding to each image the dataset also includes labels for each image having values from 0-9. The main statistics about the dataset include the minimum and maximum pixel values, mean and standard deviation of pixel values, as well as the class distribution. The minimum pixel value is 0, the maximum is 255, the mean is around 112, and the standard deviation is around 70. The class distribution is roughly uniform, with each digit class accounting for approximately 10% of the samples.

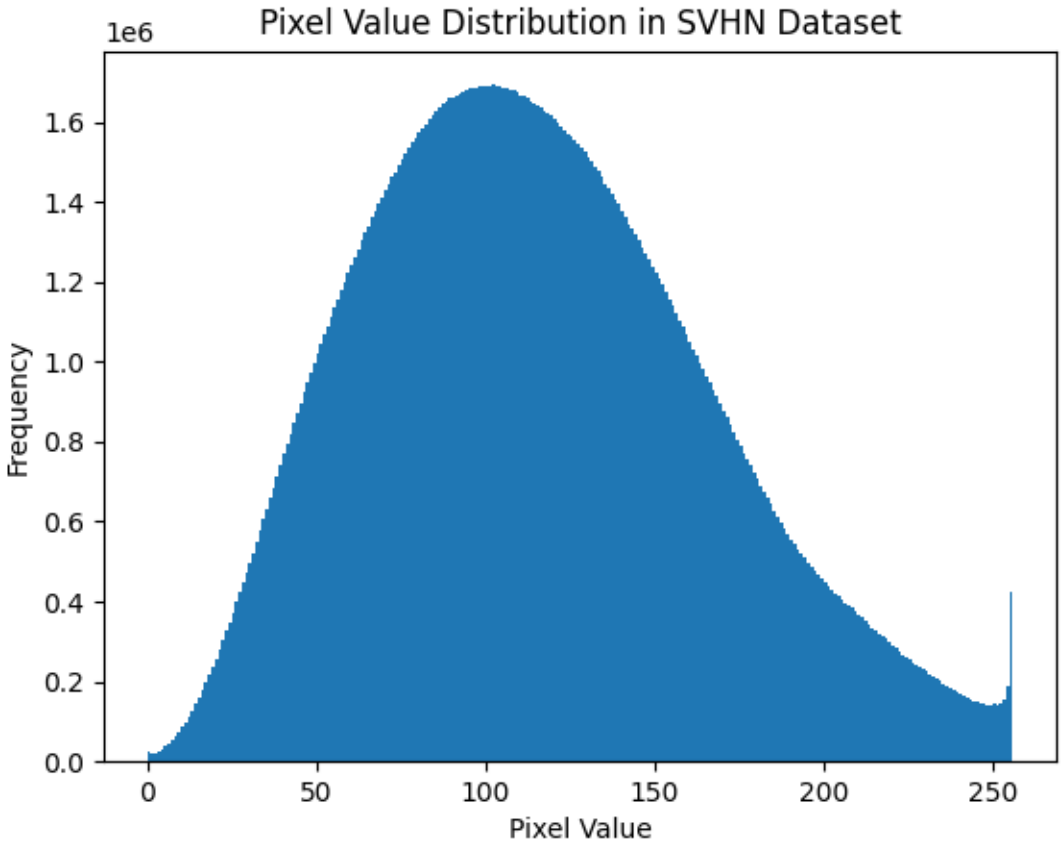
Visualization Graphs:



The above Figure is a bar plot of distribution of the classes in the training set, which shows the count of the number of samples on the y-axis and the class labels on the x-axis. It is clear from the figure that label 1 has the highest number of instances in the training set followed by label 2 and label 3.



The above figure is a 5X5 graph plotting random images along with their labels from the dataset to get a sense of what the images look like.



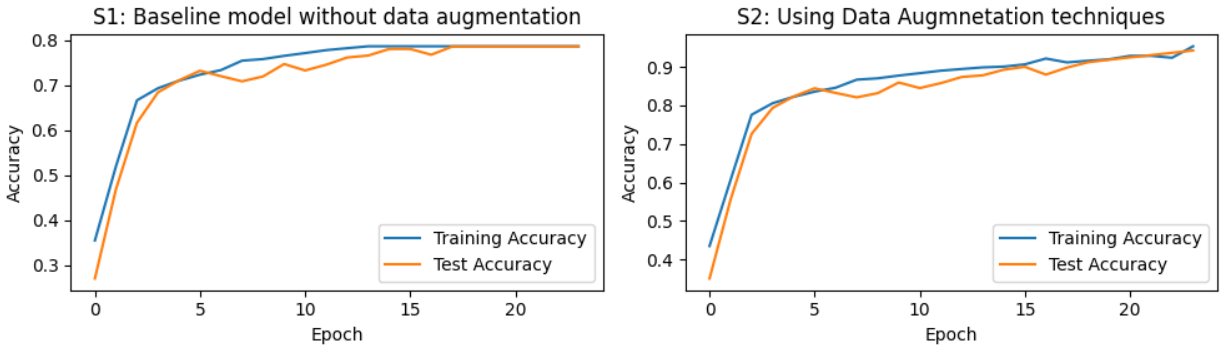
The above figure is a pixel value distribution histogram plot which shows how the pixel values are distributed across the dataset. We can see from the figure that the pixel values ranging from 80-125 have the highest frequency of occurrence.

2) The shape of the images and the number of labels in both of the datasets were different due to which we had to specify the input shape to be 64X64 for the first layer along with output shape to be 10 for our last softmax layer as in SVHN dataset we have 10 different classes.

3) Following is the list of data augmentation techniques that we used:

   a. Resizing
   b. Random Cropping
   c. Random Horizontal Flipping
   d. Random Rotation
   e. Color Jitter

To enhance our model's ability to generalize well, we incorporated diverse data augmentation techniques during the training phase. These techniques allowed our model to be trained on a wide range of input data types, thereby improving its robustness to variations and increasing its accuracy.



4)

The figure presented above showcases a comparison between two different setups - one without the implementation of any data augmentation technique and the other with data augmentation techniques. It is evident from the graph that the accuracy of the test set is significantly higher when data augmentation techniques are utilized, as opposed to when they are not used