Cat, Dog and Food Images Classification and Google SVHN Identification

Madhu Babu Sikha, Zeming Zhang
msikha@buffalo.edu, zemingzh@bufalo.edu

Abstract

In this report, we discuss the modeling of fully connected neural networks (NN) and convolutional neural networks (CNN). The initial segment involves conducting data analysis and developing a fundamental NN. The subsequent part entails understanding how to optimize and enhance the NN by employing various techniques. In the third section, we will construct a basic CNN, followed by applying optimization and data augmentation techniques in the fourth segment. We discuss the implementation of AlexNet CNN architecture for doing image classification task. We also discuss optimization of the AlexNet architecture to improve the accuracy by using Batch Normalization, Learning Rate Scheduler and Early stopping techniques. The same architecture is then used to identify Street View House Numbers (SVHN) in the next part.

1 Implementing & Improving AlexNet

In this section, we discuss details about the dataset, AlexNet architecture and its implementation for image classification and finally hyper-parameter tuning for improving the accuracy.

1.1 Details about the Dataset

The CNN dataset is a collection of 30,000 color images, where each image is of size 64x64 pixels. The images are divided into three categories: dogs, food, and vehicles. Each category contains 10,000 images, which makes the dataset balanced in terms of the number of examples per class.

These images are at different scales, colors, angles, etc., it is a very good data to train any CNN model, because the images are real-world images and hence the model can work for any test image of these categories.

Table 1: Details of CNN Dataset

Category	Number of Images	Remarks
Dogs	10,000	Different breeds of dogs, in various poses, scales, sizes, orientations, etc.
Food	10,000	Varieties of Food like burger, pizza, etc.
Vehicles	10,000	Cars of different colors, sizes, angles, etc.

1.2 Train, Validation and Test sets

For model training, validation after each epoch and final testing of our model, we have splitted the original CNN dataset into three parts namely Train, Validaiton and Test sets in 70%, 20% and 10% ratio respectively.

Table 2: CNN Dataset split for model training, validation and testing

Dataset	Split Ratio	Total Images
Train	70%	21000
Validation	20%	6000
Test	10%	3000
Total Images		30000

1.3 Sample Images Visualization

Two images from each class of CNN dataset are taken and calculated their histogram, scatter plot and heatmapt and is plotted in Figure 1.

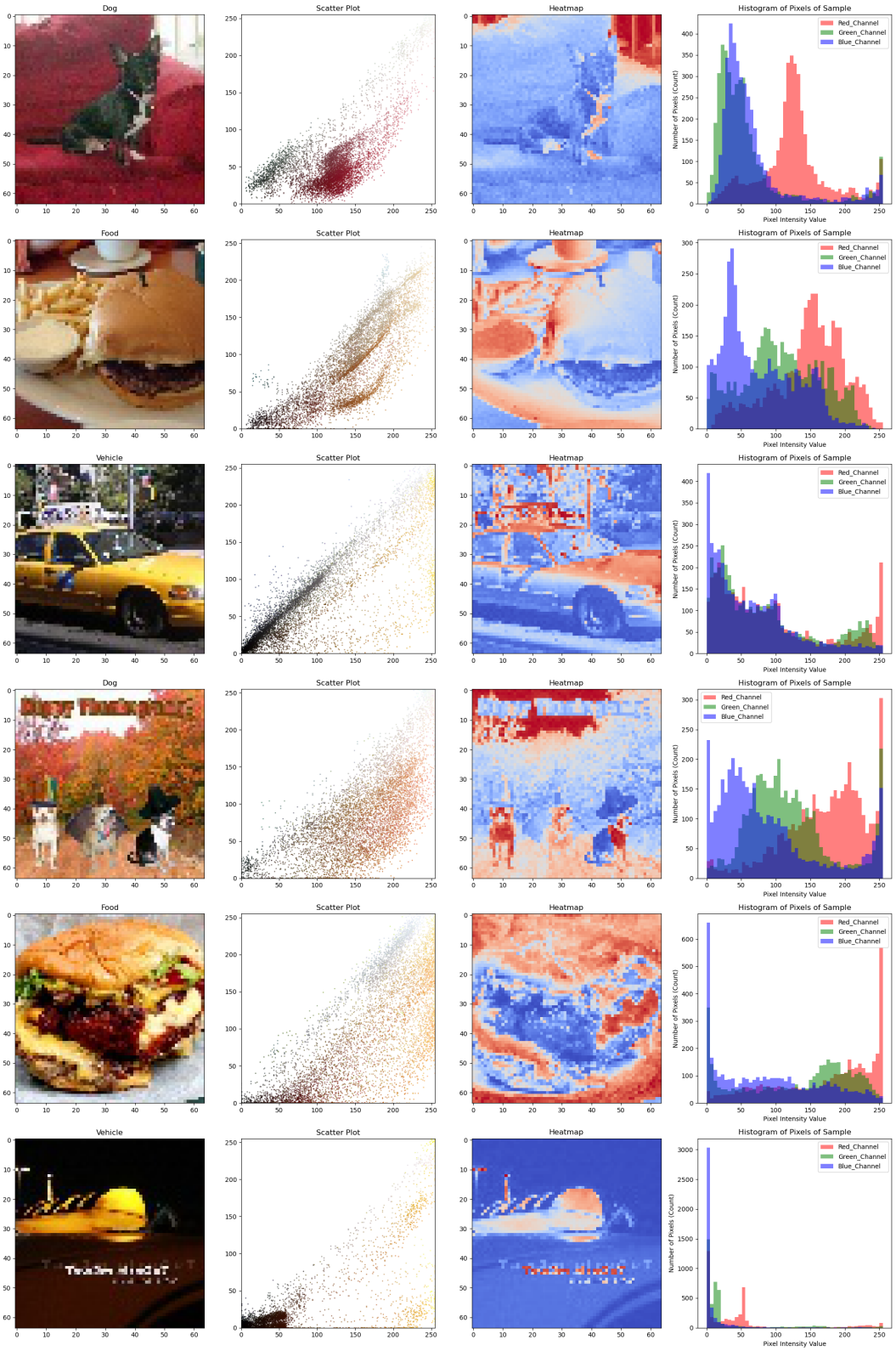


Figure 1: Sample Images of CNN dataset and their Histogram, Scatter plot and Heatmap

A scatter plot shows the relationship between the intensity values of two different color channels. By plotting the intensities of two color channels against each other, we can see if there are any patterns or relationships between them. For example, if the scatter plot shows a linear relationship between the red and green channels, this could indicate that the image contains a lot of yellow or orange colors. For food images of burger, pizza, etc., the yellow color spikes up and hence we can see that there are yellow color points in the scatter plot. If we see histogram of cars, different channels will have pixels in the same range, because usually cars color is uniform. A heatmap shows the spatial distribution of pixel intensities in an image. It can help you identify areas of high or low contrast and see how different parts of the image relate to each other. By comparing the heatmaps of different images, we can see how they differ in terms of contrast and spatial patterns.

In general, the following scatter plot trend can be observed from the CNN dataset, unless the focus is more on things other than those objects. The scatter plot for the dogs dataset shows a fairly uniform distribution of points across the plot. There doesn't appear to be any discernible pattern or clustering in the data. The scatter plot for the food dataset shows a slight clustering of points towards the center of the plot. This could be indicative of a common color or texture among the images in the dataset. The scatter plot for the cars dataset shows two distinct clusters of points. This suggests that there may be two subgroups of images in the dataset that have different features, such as color or shape.

The CNN dataset typically exhibits the following trend in the heatmap, unless the emphasis is on something other than those specific objects. The heatmap shows that the head region of the dog is the most important region for classification. This makes sense, as dogs are often identified by their facial features. The heatmap shows that the central region of the food image is the most important for classification, likely because it contains the main ingredients and visually represents the dish. The heatmap shows that the front of the car, including the headlights and grille, is the most important region for classification. This is likely because the front of a car is one of the most distinctive and recognizable parts of the vehicle.

Unless the focus is mainly on objects other than those included in the cnn dataset, the histogram trend can generally be observed. For the dogs images, we can observe a significant peak in the range of pixel intensities between 100-150, which indicates that the images have a lot of medium to high intensity pixels. This could be due to the fact that dogs have fur that tends to be darker in color, which increases the pixel intensities in those areas. In contrast, the food images have a more uniform distribution of pixel intensities, with no clear peaks or troughs. This could be because food is often photographed in evenly lit environments to bring out the colors and textures of the food, resulting in a more evenly distributed pixel intensity range. The cars images also show a clear peak in pixel intensities around the 100-150 range, similar to the dogs images. This could be due to the reflective surfaces of cars, which tend to have bright spots that increase the pixel intensities in those areas.

1.4 Building AlexNet Model

Alex et.al. [7] proposed a CNN architecture namely AlexNet for Imagenet classificaiton challenge. We have adopted the same architecture here for performing image classification, with two major changes in the output layer of the original architecture. It is discussed below: (i) the output layer of original AlexNet has 1000 neurons because it was originally designed for Imagenet challenge where they have to classify images of 1000 categories. But, in our dataset, we have only three types of images and hence we have changed the number of neurons in the output layer to 3.

Image upscaling to match the input layer size of AlexNet architecture: The input layer of original AlexNet expects images of size 224x224, while the images in our cnn dataset are of size 64x64. So, we have upscaled the image size from 64x64 to 224x224, to make them compatible for model training.

These are the two major changes we made to the original AlexNet architecture for performing our image classification task. The summary of AlexNet architecture that we have implemented is shown in Figure 2, which lists out the name of the layer, shape of the out feature map and number of parameters of each layer.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [-1, 96, 54, 54]          34,944
              ReLU-2          [-1, 96, 54, 54]               0
         MaxPool2d-3          [-1, 96, 26, 26]               0
 LocalResponseNorm-4          [-1, 96, 26, 26]               0
            Conv2d-5         [-1, 256, 26, 26]         614,656
              ReLU-6         [-1, 256, 26, 26]               0
         MaxPool2d-7         [-1, 256, 12, 12]               0
 LocalResponseNorm-8         [-1, 256, 12, 12]               0
            Conv2d-9         [-1, 384, 12, 12]         885,120
             ReLU-10         [-1, 384, 12, 12]               0
           Conv2d-11         [-1, 384, 12, 12]       1,327,488
             ReLU-12         [-1, 384, 12, 12]               0
           Conv2d-13         [-1, 256, 12, 12]         884,992
             ReLU-14         [-1, 256, 12, 12]               0
        MaxPool2d-15           [-1, 256, 5, 5]               0
AdaptiveAvgPool2d-16           [-1, 256, 6, 6]               0
          Dropout-17                [-1, 9216]               0
           Linear-18                [-1, 4096]      37,752,832
             ReLU-19                [-1, 4096]               0
          Dropout-20                [-1, 4096]               0
           Linear-21                [-1, 4096]      16,781,312
             ReLU-22                [-1, 4096]               0
           Linear-23                   [-1, 3]          12,291
================================================================
Total params: 58,293,635
Trainable params: 58,293,635
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.57
Forward/backward pass size (MB): 11.06
Params size (MB): 222.37
Estimated Total Size (MB): 234.01
----------------------------------------------------------------
```

Figure 2: AlexNet Model Architecture

1.5 Results of AlexNet Model

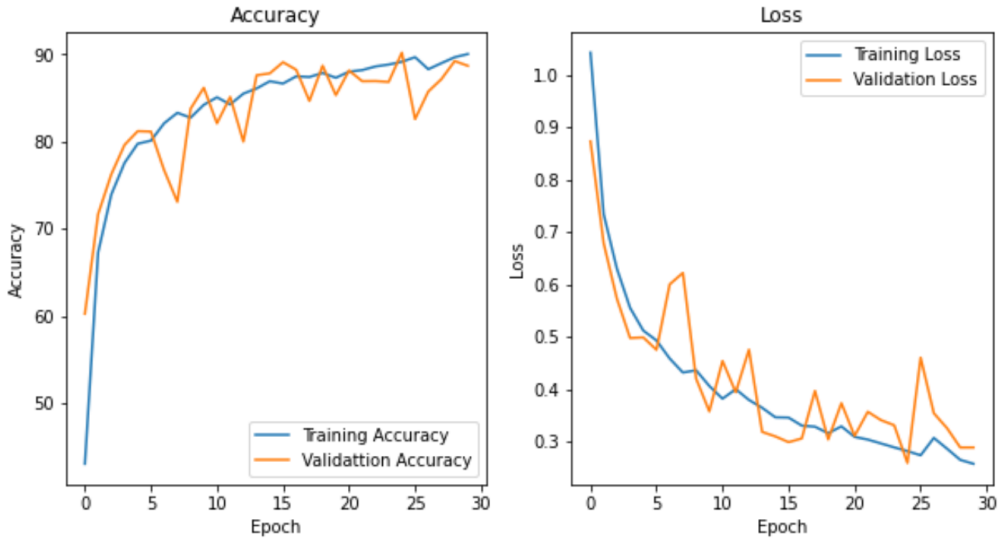
As mentioned earlier, 70% of the data is used for training the AlexNet model. After each epoch, we validate the performance of the model on the Validation set (which is 20% of cnn dataset). After training the model for predefined number of epochs, we finally test the performance of the mode on the Test set (10% of the original cnn dataset).

The parameters set for AlexNet model are given in Table 3.

Table 3: Hyperparameters for AlexNet Model (Step3 of Part-III)

Parameter	**Value**
Number of Epochs	30
Batch size	128
optimizer	SGD
Learning rate	0.01
Momentum	0.9
Activation function	ReLU

For each epoch, we recorded the average training loss and accuracy, also, the validation set loss and accuracy after every epoch is also recorded. After training phase, we plotted the training, validation loss against number of epochs as Shown in Figure 3. From this plot, we can observe that as the number of epochs increases the training loss and validation loss decreases (though there are some fluctuations in the validation loss, the overall trend is downwards only). Similarly, we have also plotted the training accuracy and validation accuracy against the number of epochs as shown in Figure 3. As the number of epochs increases, the training and validation accuracy also increases.




```
Test Loss: 0.25 -Test Accuracy: 90.73%
```

Figure 3: Loss and Accuracy vs. Epochs for Train and Validation Sets for AlexNet Model

Finally, after the training phase, we evaluated the trained (original) AlexNet model on the test set and got a train loss of 0.25, while the Train Accuracy of 90.73%.

1.6 Improved AlexNet Model

As discussed in previous subsection, the original AlexNet has given a Test accuracy of 90.73%. To further improve its performance, we have implemented three techniques namely: (i) Batch Normalization, (ii) Learning Rate Scheduler and (iii) Early Stopping.

All these three techniques are discussed in detain in Section 2.6 Optimization Methods.

Batch normalization is implemented to the original AlexNet model discussed in section 3.4. After each convolution layer in the original AlexNet architecture, a batch normalization layer is added. The batch normalization layer operates on the output of a layer, normalizing it by subtracting the mean and dividing by the standard deviation of the batch. This helps to stabilize the network by reducing internal covariate shift, which can improve training speed and performance. In other words, the batch

normalization layer can be added to any neural network layer with any number of neurons. This has helped in improving the accuracy of our model.

Also, we have implemented the learning rate scheduler to decrease the learning rate of the optimizer. There are many ways of implementing learning rate scheduler, we have reduced the learning rate for each epoch as a function of the current epoch, as given in Table 4. This type of learning rate scheduler is called Exponential decay. This can help to improve the performance of the model by allowing it to make larger updates to the parameters early in training when the gradients are larger, and smaller updates later in training when the gradients are smaller.

For early stopping we set a patience value of 8, i.e., if the validation loss of the current epoch increases than that of the previous epoch 8 number of times, then the training will be stopped. Since we ran the model for 32 epochs and there is no significant over-fitting of the model, early-stopping did not stop the training phase. During our training phase, we have observed the nature of the dataset and the model and did not see the validation loss in the current epoch increasing than the previous epoch. So, early stopping didn't help us in stopping the training early and hence there is no improvement in the accuracy because of early-stopping.

The parameters of Modified AlexNet are given in Table 4.

Table 4: Hyperparameters for Modified AlexNet Model (Step4 of Part-III)

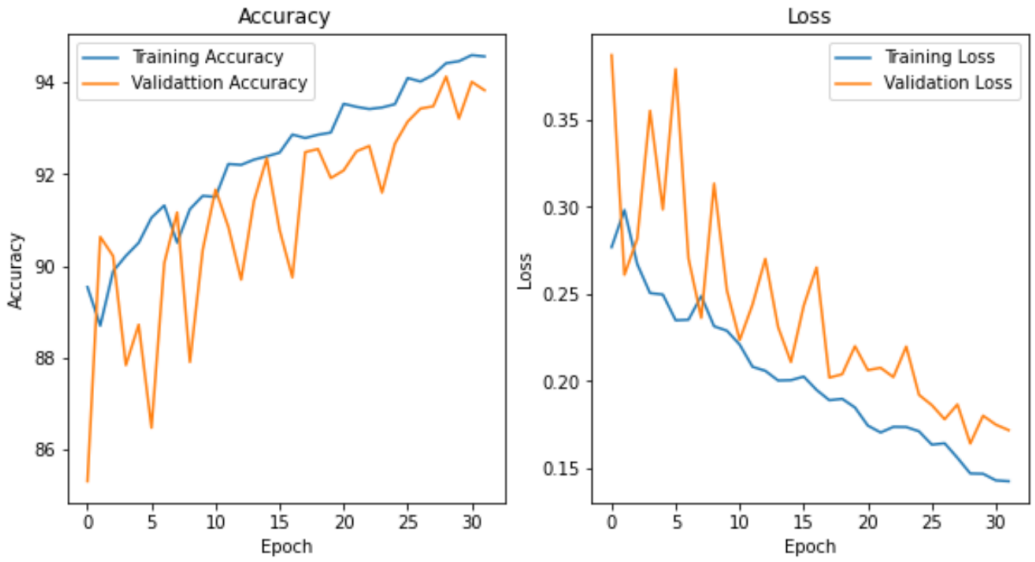
Parameter	**Value**
Number of Epochs	32
Batch size	128
optimizer	SGD
Learning rate	0.01
Momentum	0.9
Activation function	ReLU
BatchNormalization	appropriate to the previous Conv Layer
Early stopping	patience of 8
Learning Rate Scheduler	initial_lr * (1-(current_epoch/total_epochs))

1.7 Results of Train, Validation and Test for Improved AlexNet Model

During each epoch, we monitored and recorded the average training loss and accuracy, as well as the validation set loss and accuracy. Subsequently, we plotted the training and validation loss against the number of epochs, as depicted in Figure 5. The plot reveals a decline in both training and validation loss as the number of epochs increases. Although the validation loss exhibits some fluctuations, the overall trend is downward. In addition, we also plotted the training accuracy and validation accuracy against the number of epochs, as presented in Figure 5. As the number of epochs increases, the training and validation accuracy also increase.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [-1, 96, 54, 54]          34,944
       BatchNorm2d-2          [-1, 96, 54, 54]             192
              ReLU-3          [-1, 96, 54, 54]               0
         MaxPool2d-4          [-1, 96, 26, 26]               0
 LocalResponseNorm-5          [-1, 96, 26, 26]               0
            Conv2d-6         [-1, 256, 26, 26]         614,656
       BatchNorm2d-7         [-1, 256, 26, 26]             512
              ReLU-8         [-1, 256, 26, 26]               0
         MaxPool2d-9         [-1, 256, 12, 12]               0
LocalResponseNorm-10         [-1, 256, 12, 12]               0
           Conv2d-11         [-1, 384, 12, 12]         885,120
      BatchNorm2d-12         [-1, 384, 12, 12]             768
             ReLU-13         [-1, 384, 12, 12]               0
           Conv2d-14         [-1, 384, 12, 12]       1,327,488
      BatchNorm2d-15         [-1, 384, 12, 12]             768
             ReLU-16         [-1, 384, 12, 12]               0
           Conv2d-17         [-1, 256, 12, 12]         884,992
      BatchNorm2d-18         [-1, 256, 12, 12]             512
             ReLU-19         [-1, 256, 12, 12]               0
        MaxPool2d-20           [-1, 256, 5, 5]               0
AdaptiveAvgPool2d-21           [-1, 256, 6, 6]               0
          Dropout-22                [-1, 9216]               0
           Linear-23                [-1, 4096]      37,752,832
             ReLU-24                [-1, 4096]               0
          Dropout-25                [-1, 4096]               0
           Linear-26                [-1, 4096]      16,781,312
             ReLU-27                [-1, 4096]               0
           Linear-28                   [-1, 3]          12,291
================================================================
Total params: 58,296,387
Trainable params: 58,296,387
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.57
Forward/backward pass size (MB): 15.64
Params size (MB): 222.38
Estimated Total Size (MB): 238.60
----------------------------------------------------------------
```

Figure 4: Improved AlexNet Model Architecture




```
Test Loss: 0.17 -Test Accuracy: 93.63%
```

Figure 5: Loss and Accuracy vs. Epochs for Train and Validation Sets for Modified AlexNet Model

Finally, after the training phase, we evaluated the trained modified and improved AlexNet model on the test set and got a train loss of 0.17, while the Train Accuracy of 93.63%, which is better than the original AlexNet model.

2 Optimizing CNN + Data Argumentation of Google SVHN Dataset

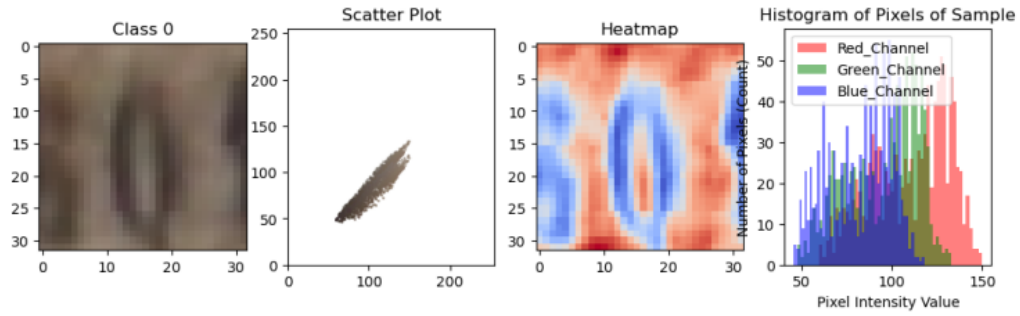
2.1 Details about the Dataset

In this part, we do image classification of Street View House Numbers (SVHN) Dataset [8] using the optimized AlexNet architecture from part-III with few modifications (discussed in next section).

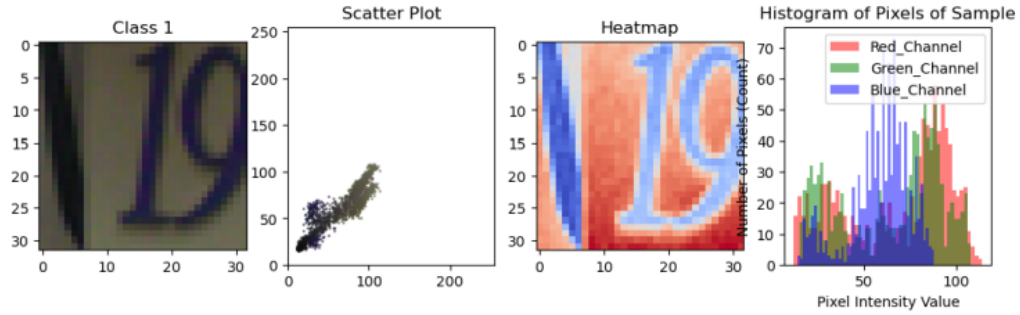
The Google Street View House Numbers (SVHN) dataset is a large dataset of digit images that was collected from house numbers visible in Google Street View images. The dataset was released in 2011 by Google and has since become a popular benchmark dataset for image classification tasks, particularly for digit recognition. We have downloaded the data from torchvision.datasets import SVHN, where the training set contains 73,257 images, the validation set (which is loaded from the 'extra' split) contains 26,032 images, and the test set (also loaded from the 'test' split) contains 26,032 images. The images are labeled with the corresponding digit, and there are 10 classes, one for each digit from 0 to 9. One of the challenges of the SVHN dataset is that the digits in the images may be overlapping or occluded, making it more difficult to accurately classify them. Additionally, the dataset contains some images that are blurry or distorted, which can further complicate the task of digit recognition.

The analysis of a sample image from each class in the google SVHN dataset is given in the Figures 6, 7 and 8.

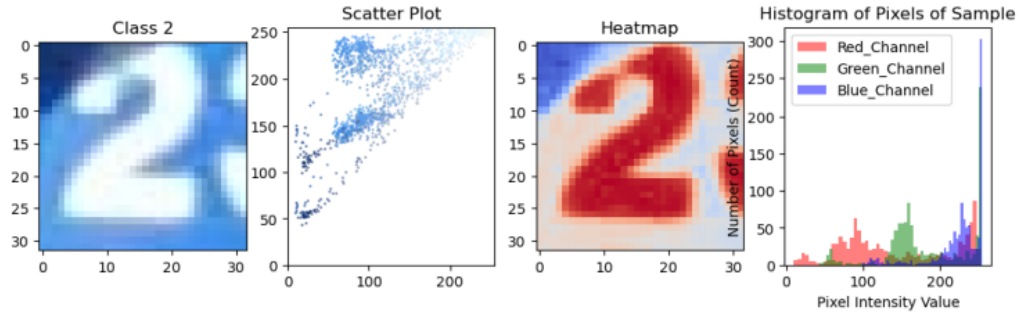
All the images are of 32x32 and hence the number of pixels is less. The scatter plot of these images



<Figure size 640x480 with 0 Axes>



<Figure size 640x480 with 0 Axes>



<Figure size 640x480 with 0 Axes>

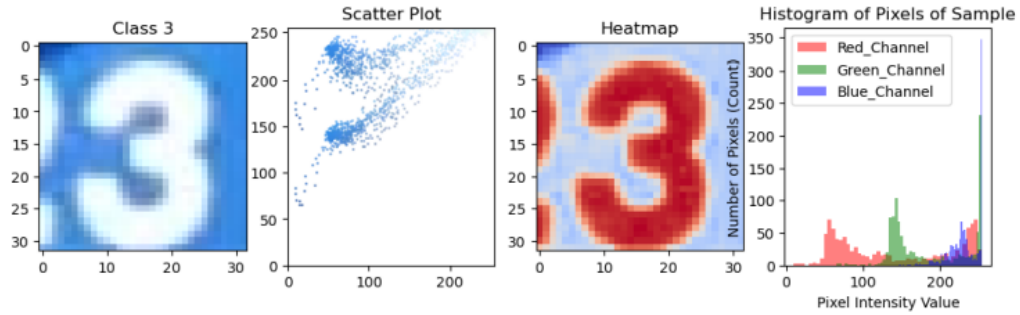
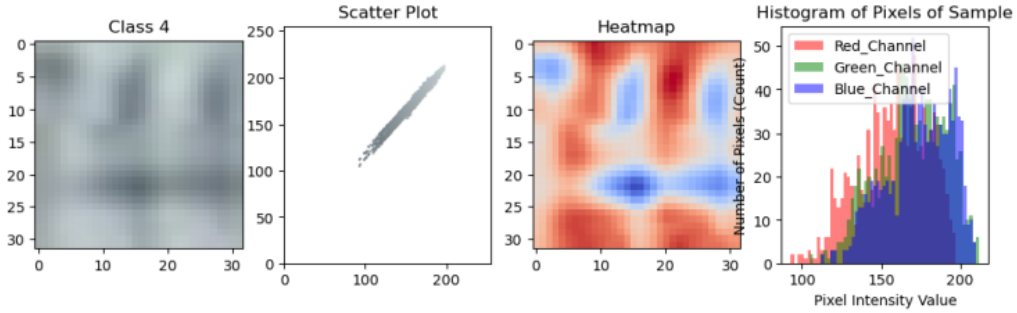
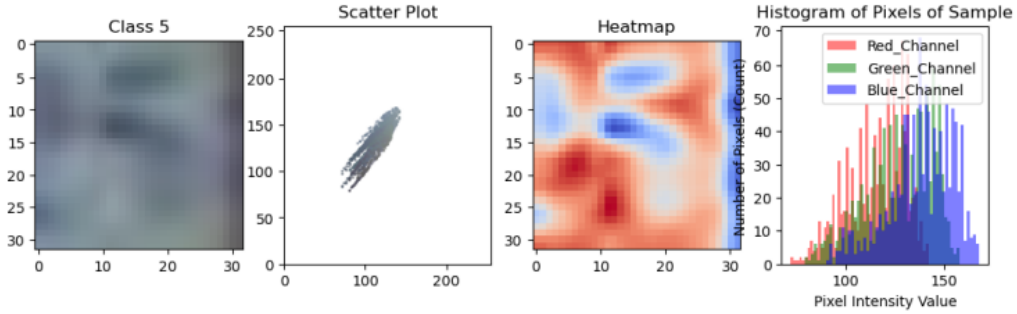


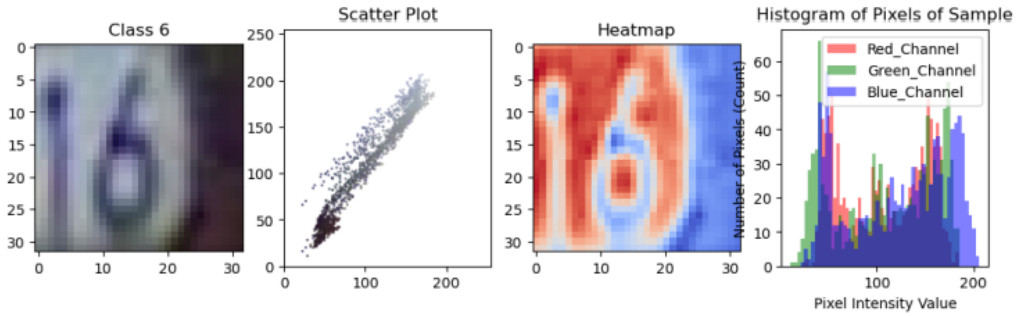
Figure 6: Google SVHN dataset digits 0, 1, 2, 3



<Figure size 640x480 with 0 Axes>



<Figure size 640x480 with 0 Axes>



<Figure size 640x480 with 0 Axes>

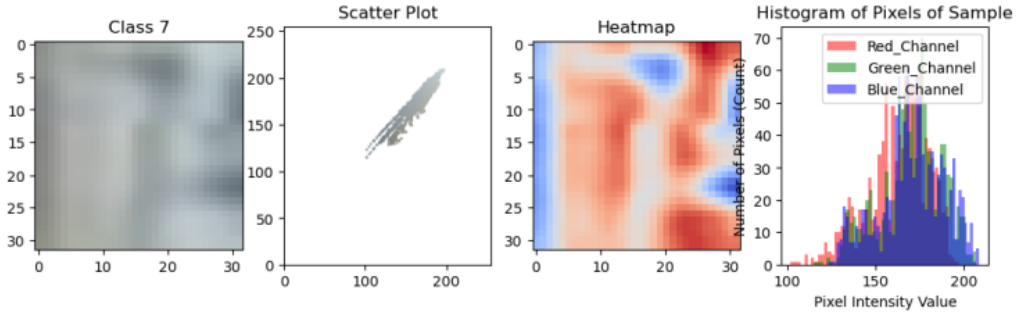


Figure 7: Google SVHN dataset digits 4,5,6,7

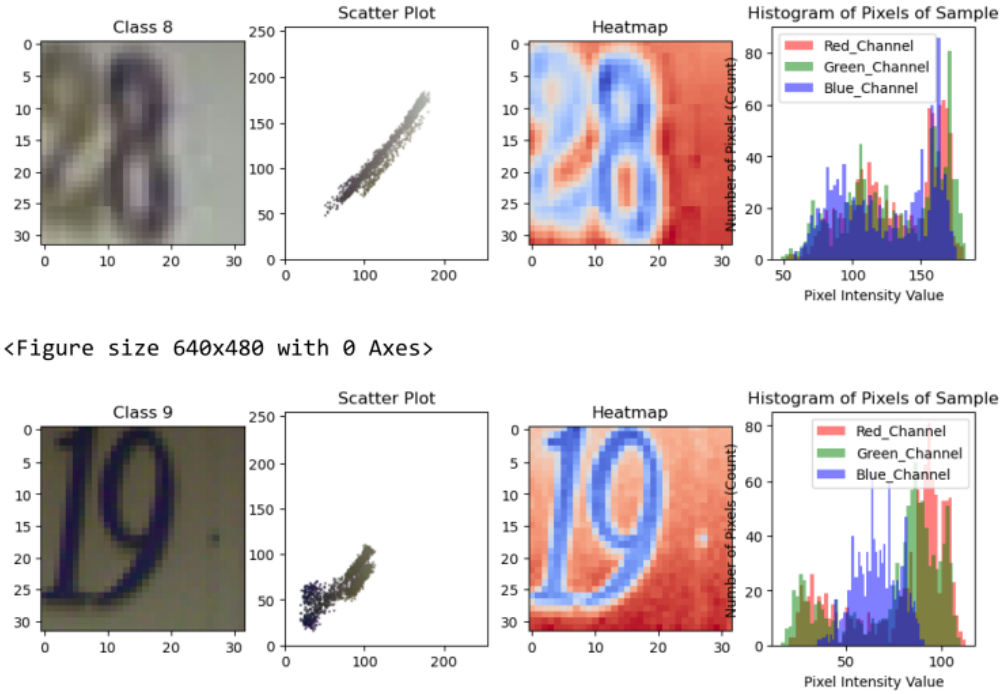


Figure 8: Google SVHN dataset digits 8, 9

We compare the histograms of the SVHN images, we might notice that the histograms of images containing house numbers with multiple digits have a wider range of pixel intensities than those containing only single digits. This could be due to the fact that images with multiple digits contain more complex backgrounds, colors, and textures.

Similarly, we notice that the histograms of images containing blurry or low-quality house numbers have less distinct peaks, indicating that the pixel intensities are more evenly distributed across the image. This could be due to the fact that blurry or low-quality images have less contrast between the foreground (house number) and background, making it harder to distinguish the individual pixels.

Analyzing histograms can provide valuable insights into the properties of images, and can be used to compare and contrast different types of images.

In the case of SVHN dataset, which contains images of house numbers, the heatmaps can help in identifying the regions of the image that contain the number and its location within the image.

For example, if we compare the heatmaps of different images of the same house number, we can observe that the regions with higher intensity values correspond to the areas where the number is present, while the background has lower intensity values. Additionally, we can also observe variations in the heatmaps for the same house number, which can be attributed to variations in the lighting conditions, image quality, and other factors.

Overall, comparing heatmaps can help in understanding the variations and patterns within a dataset, which can be useful for developing better models and improving the accuracy of image classification tasks.

The scatter plots of the SVHN dataset show the distribution of the RGB color values for each image. By looking at the plots, we can observe that the distribution of color values varies across different images. Some images have a uniform distribution, while others have more variation in their color values. Additionally, some images have a dominant color that stands out in the scatter plot.

These differences in the scatter plots can provide insights into the characteristics of the images. For example, images with a dominant color could be associated with certain types of objects, such as

traffic signs or buildings. By examining the scatter plots, we can identify these patterns and potentially use them to improve our image recognition models.

2.2 Improved AlexNet Model from Part-III and Modifications

In this part, to solve the Google SVHN challenge, We have adopted the same AlexNet architecture that we have optimized in part-III, but with one major change in the output layer of the original architecture. It is discussed below: (i) the output layer of original AlexNet has 1000 neurons because it was originally designed for Imagenet challenge where they have to classify images of 1000 categories. But, in our Google SVHN dataset, we have only 10 types of images and hence we have changed the number of neurons in the output layer to 10.

Image upscaling to match the input layer size of AlexNet architecture: The input layer of original AlexNet expects images of size 224x224, while the images in Google SVHN are of 32x32 size. So, we have upscaled the image size from 32x32 to 224x224, to make them compatible for model training.

2.3 Data Augmentation Methods

Data augmentation is a technique used in machine learning to increase the amount of training data available for a model by creating modified versions of the original data. The goal of data augmentation is to improve the generalization of the model by exposing it to more diverse examples and reducing overfitting.

In this work, to generate a new images, from the original dataset, we have used the following techniques: rotation, horizontal flip, vertical flip, color jitter. Since we have increased the number of training images, the training time also increased. But the model has got exposed to a variety of images, i.e., the diversity of the images is increased. Also, the size of the dataset increased as well, which helped us in reducing the overfitting and hence increase in the accuracy on the test.

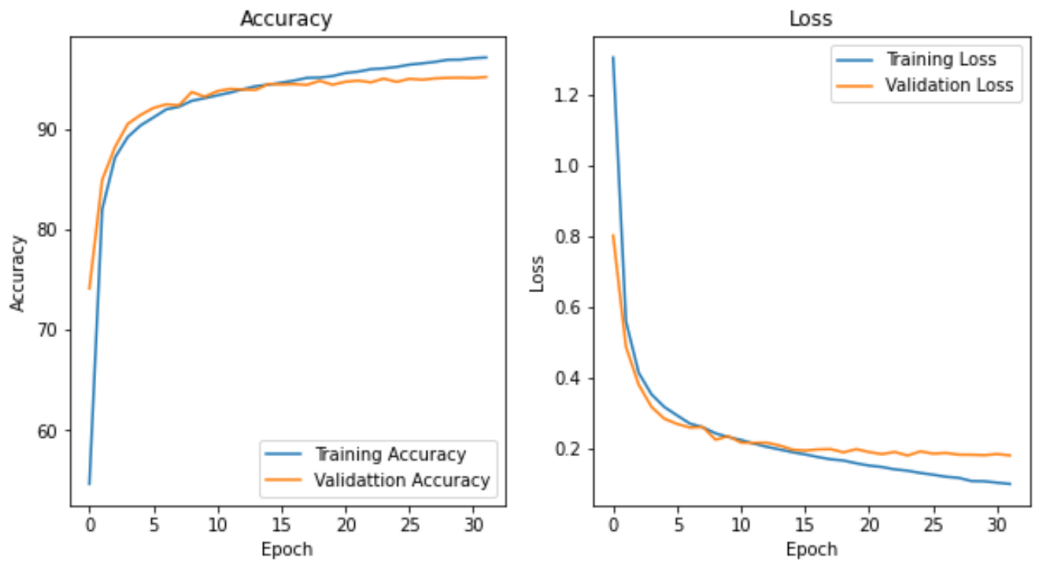
2.4 Results of Train, Validation and Test of SVHN Dataset using Improved AlexNet Model

We train the improved AlexNet architecture with output layer changes and input image upscaling. The parameters of the improved AlexNet model are kept same as part-III improved model which are given in Table 4.

As discussed earlier, we have noted the training, validation loss and accuracy for every epoch and plotted the results in Figure 10. This plot shows that as the number of epochs increases, both training and validation loss decrease. In addition, we can see that the training and validation accuracy also increase as the number of epochs increases. This suggests that the model is learning and improving over time. The results are visualized in Figure 10.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1           [-1, 96, 54, 54]          34,944
       BatchNorm2d-2           [-1, 96, 54, 54]             192
              ReLU-3           [-1, 96, 54, 54]               0
         MaxPool2d-4           [-1, 96, 26, 26]               0
 LocalResponseNorm-5           [-1, 96, 26, 26]               0
            Conv2d-6          [-1, 256, 26, 26]         614,656
       BatchNorm2d-7          [-1, 256, 26, 26]             512
              ReLU-8          [-1, 256, 26, 26]               0
         MaxPool2d-9          [-1, 256, 12, 12]               0
LocalResponseNorm-10          [-1, 256, 12, 12]               0
           Conv2d-11          [-1, 384, 12, 12]         885,120
      BatchNorm2d-12          [-1, 384, 12, 12]             768
             ReLU-13          [-1, 384, 12, 12]               0
           Conv2d-14          [-1, 384, 12, 12]       1,327,488
      BatchNorm2d-15          [-1, 384, 12, 12]             768
             ReLU-16          [-1, 384, 12, 12]               0
           Conv2d-17          [-1, 256, 12, 12]         884,992
      BatchNorm2d-18          [-1, 256, 12, 12]             512
             ReLU-19          [-1, 256, 12, 12]               0
        MaxPool2d-20            [-1, 256, 5, 5]               0
AdaptiveAvgPool2d-21            [-1, 256, 6, 6]               0
          Dropout-22                 [-1, 9216]               0
           Linear-23                 [-1, 4096]      37,752,832
             ReLU-24                 [-1, 4096]               0
          Dropout-25                 [-1, 4096]               0
           Linear-26                 [-1, 4096]      16,781,312
             ReLU-27                 [-1, 4096]               0
           Linear-28                   [-1, 10]          40,970
================================================================
Total params: 58,325,066
Trainable params: 58,325,066
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.57
Forward/backward pass size (MB): 15.64
Params size (MB): 222.49
Estimated Total Size (MB): 238.71
----------------------------------------------------------------
```

Figure 9: Improved AlexNet Model Architecture for Google SVHN dataset




```
Test Loss: 0.18 -Test Accuracy: 95.31%
```

Figure 10: Loss and Accuracy vs. Epochs for Train and Validation Sets for Google SVHN Dataset

Finally, after the training phase, we evaluated the trained modified and improved AlexNet model on the test set of SVHN dataset and got a train loss of 0.18, while the Train Accuracy of 95.31%.

2.5 Different Setups for Part-IV

The optimized AlexNet model from part-III has given us an accuracy of 95.31% as mentioned above. Though we have tried two more variants of AlexNet, we did not see much improvement in the accuracy.

The first variant is a deeper version of the network with two additional convolutional layers, 384 and 256 filters respectively, and reduced filters in the first layer to 64. We also removed local response normalization layers and did not add any additional regularization techniques.

The second variant is an improved version of AlexNet, where we increased the number of filters, added batch normalization after every convolutional layer, and dropout after every fully connected layer. We replaced nn.BatchNorm2d with nn.BatchNorm1d after the fully connected layers, omitted local response normalization layers, and did not add data augmentation techniques.

3 Conclusion

In this assignment, we have explored and implemented fully connected neural networks (NN) and convolutional neural networks (CNN) in four different parts.

We implemented AlexNet architecture to classify images in CNN dataset of 3 classes, namely dogs, food and cars. We observed the impact of varying the number of convolutional layers, filters and other hyperparameters on the accuracy of the model. We also applied optimization techniques on the AlexNet to improve its performance further.

In the next part, we applied optimization techniques such as data augmentation on Google SVHN dataset and classified them using the optimized AlexNet from third part. We observed the effect of different data augmentation techniques on the accuracy of the model.

Overall, this work has allowed us to gain hands-on experience in building and optimizing different types of neural networks for image classification tasks. We have also learned about important

techniques such as early stopping, learning rate scheduler, data augmentation, and batch normalization, which can help improve the performance of neural networks.

References

[1] Pytorch documentation. https://pytorch.org/docs/stable/index.html. Last accessed on 03-30-2023.

[2] Numpy documentation. https://numpy.org/doc/stable/. Last accessed on 03-31-2023.

[3] Pandas documentation. https://pandas.pydata.org/docs/index.html. Last accessed on 03-31-2023.

[4] Lecture slides of Intro to Machine Learning course by Dr. Alina Vereshchaka, Dept. of CSE, UB.

[5] Machine Learning Mastery. https://machinelearningmastery.com/develop-your-first-neural-network-with-pytorch-step-by-step/. Last accessed on 03-31-2023.

[6] Stack Overflow. https://stackoverflow.com/. Last accessed on 03-31-2023.

[7] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60.6 (2017): 84-90.

[8] The Street View House Numbers (SVHN) Dataset. Last accessed on 04-13-2023.