DS5220 Supervised Machine Learning

Final Project

By:

Pranav Vishwanath (NUID:002766766)
Raghav Gali (NUID:002271909)

Methodologies

We have utilized the Jupyter Notebook Environment for Python to perform our analysis.We have opted for the following models for our analysis of the CIAR-100 Dataset:-

1. Support Vector Machine Without Kernel (SVM)
2. Support Vector Machine With Radial Bias Function Kernel (Kernel SVM)
3. Logistic Regression
4. Convoluted Neural Networks

We will be explaining our results of each model in detail along with our methodologies in this report.

Data Processing

Firstly, to get a feel for the dataset, we do some eda and see a few sample images and their dimensions - 32X32X3

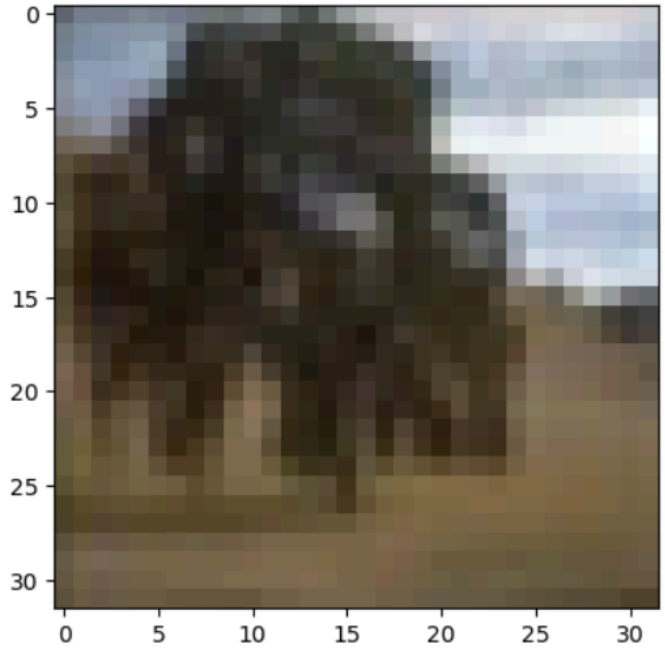


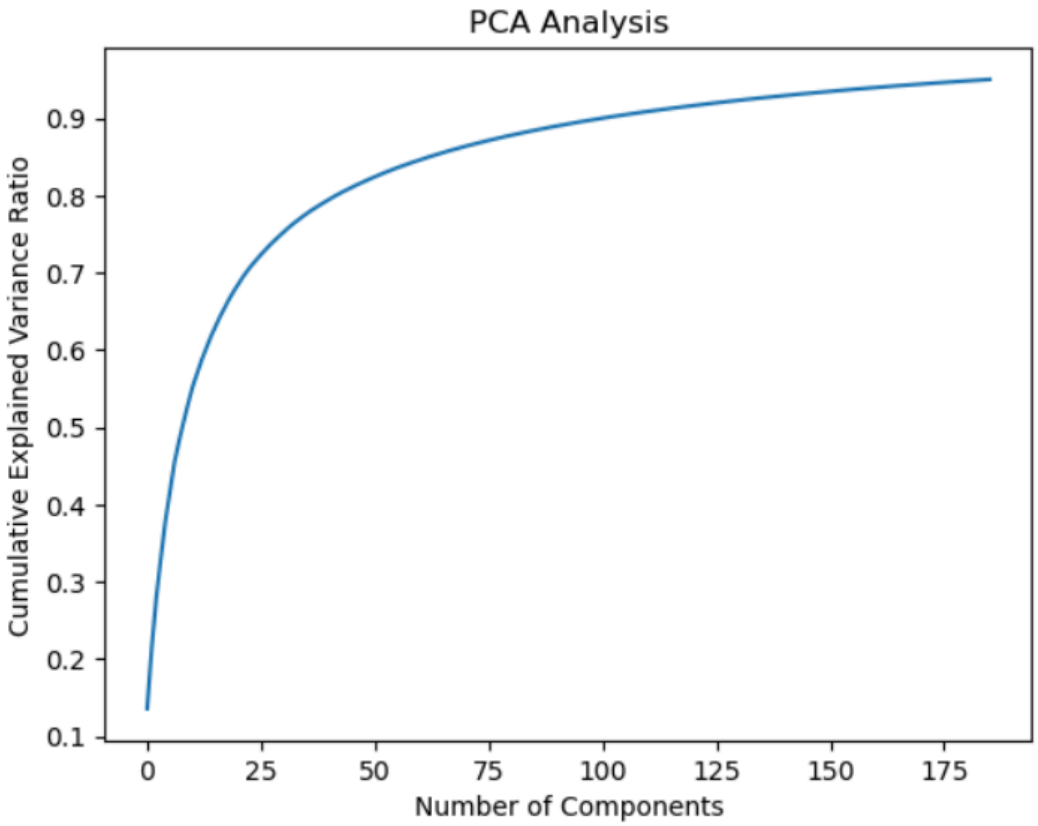
Image Label: [52]

To tackle processing the CIFAR-100 dataset for SVMs and Logistic Regression, we first need to pre-process the dataset. The images of 32X32X3 are normalized into 0.5X0.5X0.5 and passed into the models.

Initial Model Performance

1. Support Vector Machine without Kernel:- This model initially achieved an accuracy of 13.42% on the cifar 100 dataset.
2. Support Vector Machine with rbf Kernel:- This model initially achieved accuracy of 26.39% on the cifar 100 dataset.
3. Logistic Regression :- This model initially achieves an accuracy of 27% on the cifar 100 dataset.
4. Convoluted Neural Networks:- This model initially achieves an accuracy of __ on the cifar 100 dataset.

PCA and Feature Extraction

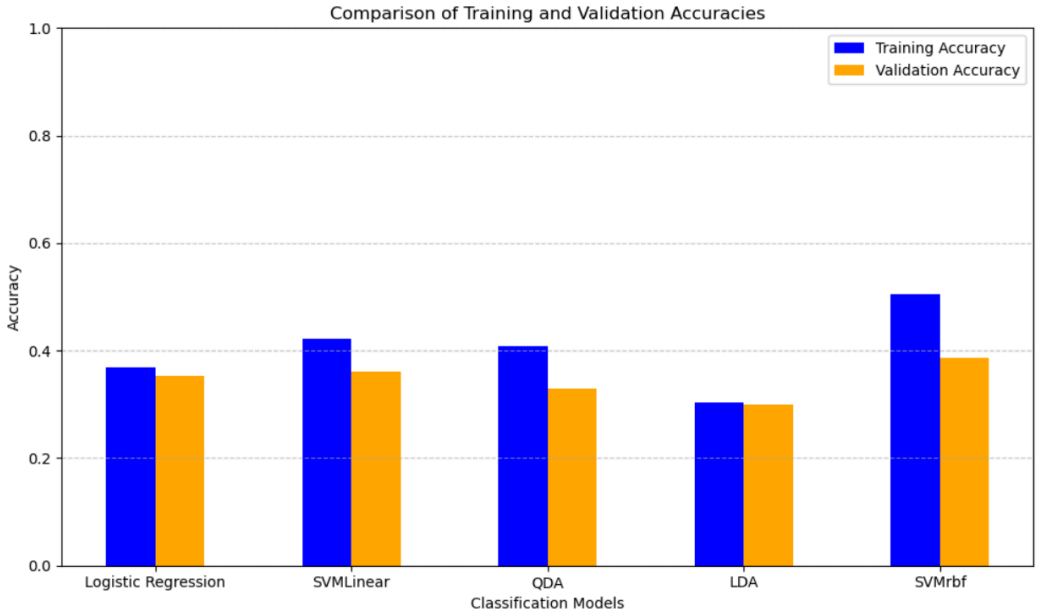


From the PCA analysis graph on the cifar-100 Dataset, we can observe that to capture 80% of the variance, around 50 components are sufficient. Hence, we will set a threshold of 0.8 for the variance in PCA and perform our analysis on that.

We perform feature extraction using the resnet model weights trained on image datasets to further improve the accuracy, there are significant improvements once these methods are performed.

We also filter out the 20 superclasses in the dataset to further improve the model performance and processing times.

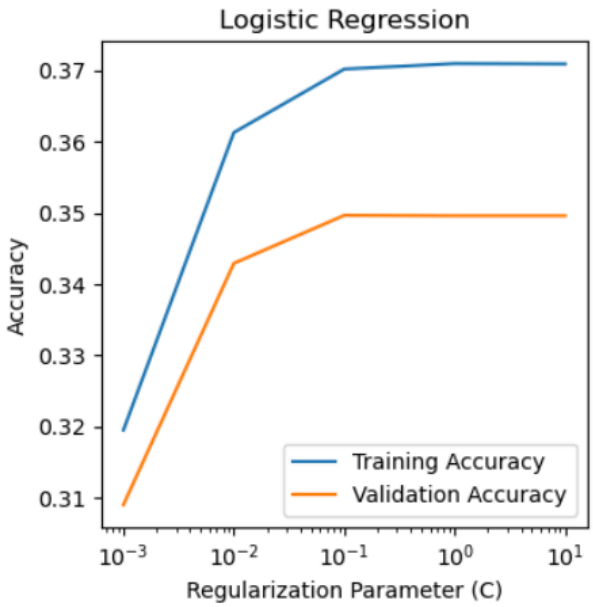
The results after these three steps are as follows:

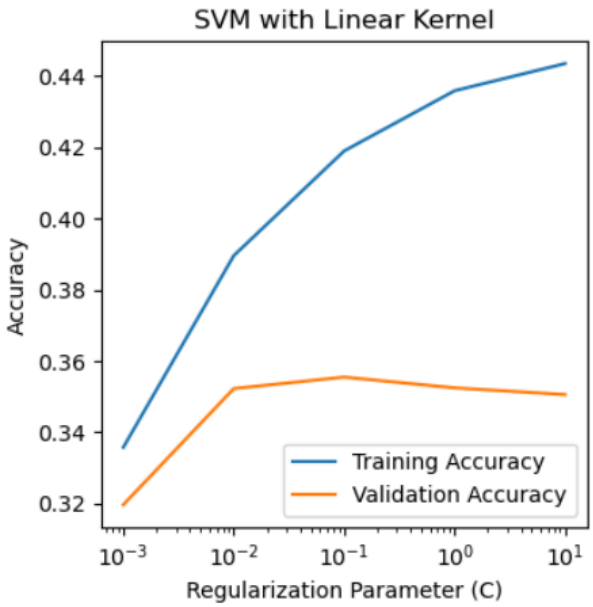


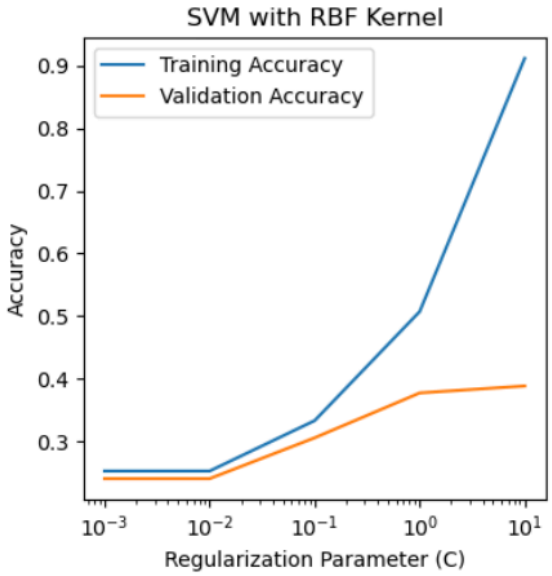
Logistic Regression, Linear SVM and rbf SVM all find significant improvements in the accuracies and just for a comparison we have included LDA and QDA too. Out of all these models, kernel SVM provides the best accuracy with around 56%.

L2 Regularization

After performing L2 regularization on the data set to further increase the accuracy, theTraining and Validation Set of the accuracies are as follows:- Note that this was performed on the entire dataset of 100 classes but not 20



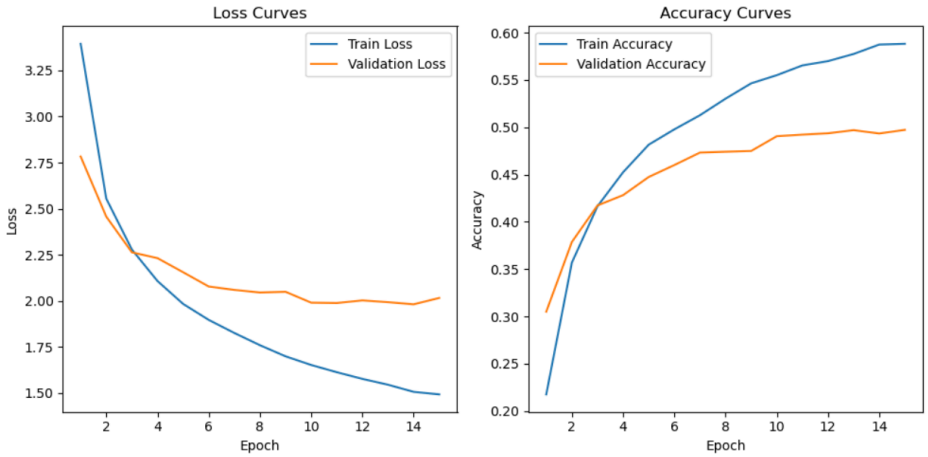




We can see that even though the training accuracy is high for kernel SVM, the validation accuracy is comparable to the other models. This could be because of an imbalance in any of these classes sets.

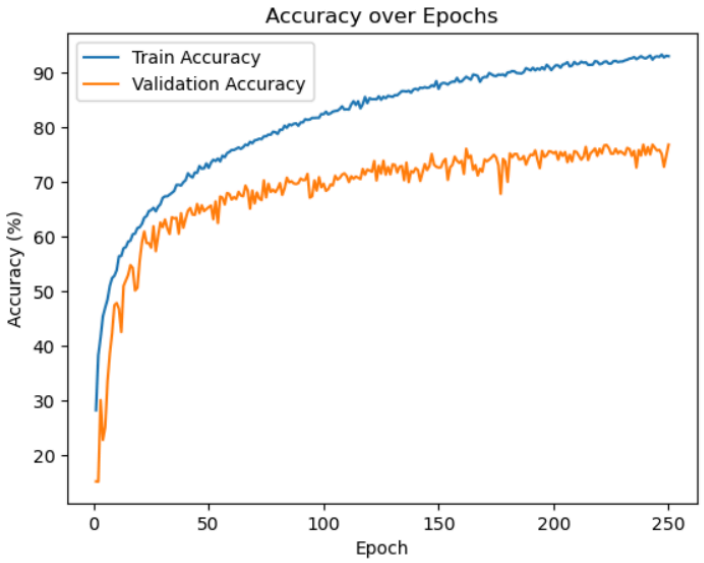
Convolutional Neural Network

- **All 100 classes without regularization:-**

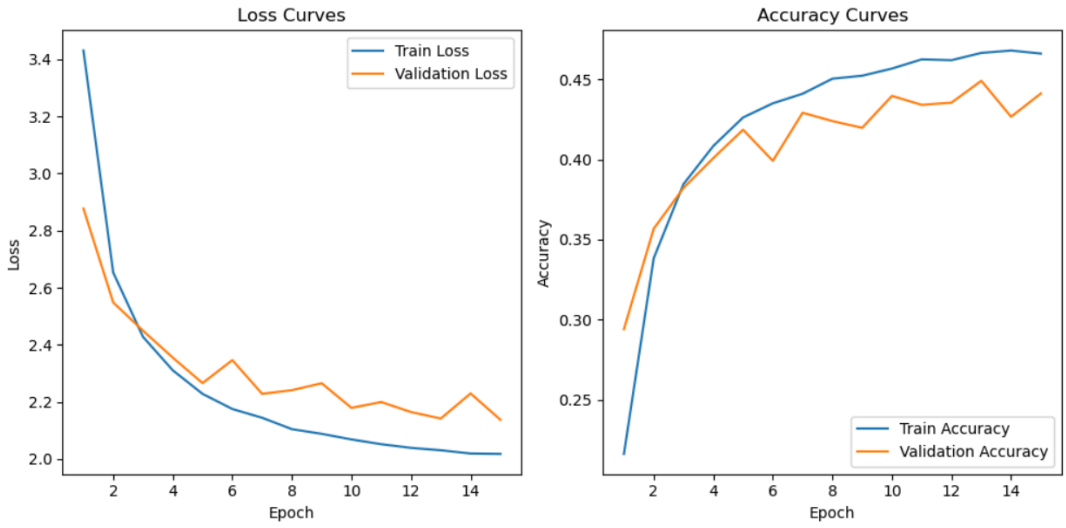


The following results were obtained by running the Neural Network on the CNN Dataset using ReLU activation functions without regularization. We obtained a maximum accuracy of 60% and the plot of the loss functions and accuracies at each epoch have been illustrated in the plot above.

- **For 20 Superclasses with Regularization: -**
  The following results were obtained by running the Neural Network on the CNN Dataset using ReLU activation functions without regularization. We obtained a maximum accuracy of 82% and the plot of the accuracies at each epoch have been illustrated in the plot below.



- **For 100 Classes with Regularization: -**



To conclude, with Regularization, we are able to achieve similar results on the cifar 100 dataset suggesting that the superclass classification is a better approach to building a model on this dataset