Data Mining

Project Report

Razi Haider

19I-1762

BS(DS)-N

Original Image vs. Segmented Image

A picture containing balloon, dark, aircraft, light

Description automatically generatedA close-up of a person's skin

Description automatically generated with medium confidence

A close up of a planet

Description automatically generated with low confidenceA close-up of a person's skin

Description automatically generated with medium confidence

K-means was used to segment images with k having the value of seven (explained latter). The size of the images was reduced to 100x100 for faster performance and to avoid overfitting.

Models:

* ANN
* CNN
* Gaussian NB
* Logistic Regression

Clustering:

Chart, line chart

Description automatically generated

I used K-elbow to check which value of k is better for clustering and the graph suggests that k should be seven. The value of k was fluctuating between eight, nine, and seven every time I used K-elbow. The clustered were then formed using seven labels which were derived by clustering images.

*Reference:* [*https://www.geeksforgeeks.org/image-segmentation-using-k-means-clustering/*](https://www.geeksforgeeks.org/image-segmentation-using-k-means-clustering/)

Performance:

Chart, bar chart

Description automatically generated

After clustering and segmenting images, I used various models on the data and recorded accuracies for all the used models.

Among traditional ML models, gaussian NB performed better than logistic regression. This is mainly due to the fact that logistic regression has parts that maps data in linear plane along with scaling data into zeros and ones.

Among NN models, the ANN and CNN models performed better in terms of accuracy with the images were not segmented. The figure also shows ANN outperforming CNN for original images. However, accuracy is not always better. Loss (error) and the time taken also matters.

Each NN model for run till twenty epochs.

Chart, bar chart

Description automatically generated

The above figure suggests that the losses for CNN models are lower as compared to ANN. This suggests that CNN is performing better than ANN. However, we also have to look at time taken by each model.

Chart

Description automatically generated with low confidence

This above graph shows that the time taken by NN models on segmented images run significantly faster than NN models on original images and there is no significant difference in terms of accuracies and losses.

Among traditional ML models, the gaussian NB runs significantly faster than logistic regression. Gaussian NB is also better in terms of accuracy.