Assignment -2
K-means Clustering and Auto Encoders (Unsupervised Learning)

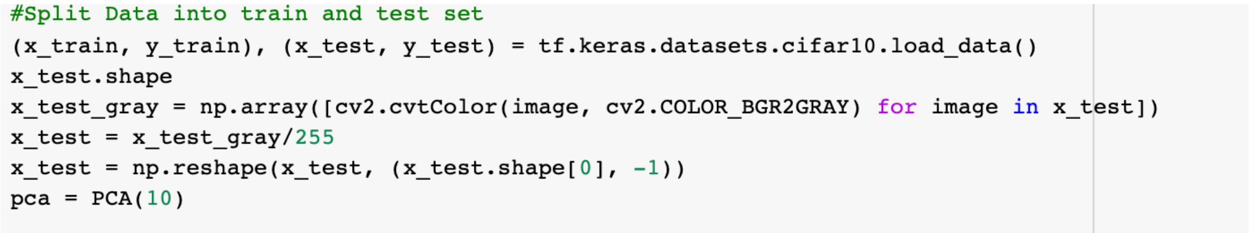
Nikhil Bandari (Person Number: 50418501)   14 November 2021

1 Assignment Overview:

The goal of the assignment is working on K-means clustering and auto encoders on Cifar 10 Data Set. In the first part of the assignment, I have generated 10 clusters on test set of Cifar of 10,000 examples. In the Second phase, I used same dataset for auto-encoders to get encoded and decoded compressed images to get reconstructed and compared with original images.

2 Data -Pre-Processing:

Each example is a 32x32 image, the test data(x_test) is converted into grayscale images using cv2 Library and reshaped to 255 images.



```
#Split Data into train and test set
(x_train, y_train), (x_test, y_test) = tf.keras.datasets.cifar10.load_data()
x_test.shape
x_test_gray = np.array([cv2.cvtColor(image, cv2.COLOR_BGR2GRAY) for image in x_test])
x_test = x_test_gray/255
x_test = np.reshape(x_test, (x_test.shape[0], -1))
pca = PCA(10)
```

3 Python Editor

I have used Google Colab IDE for implementation and Comparing results.

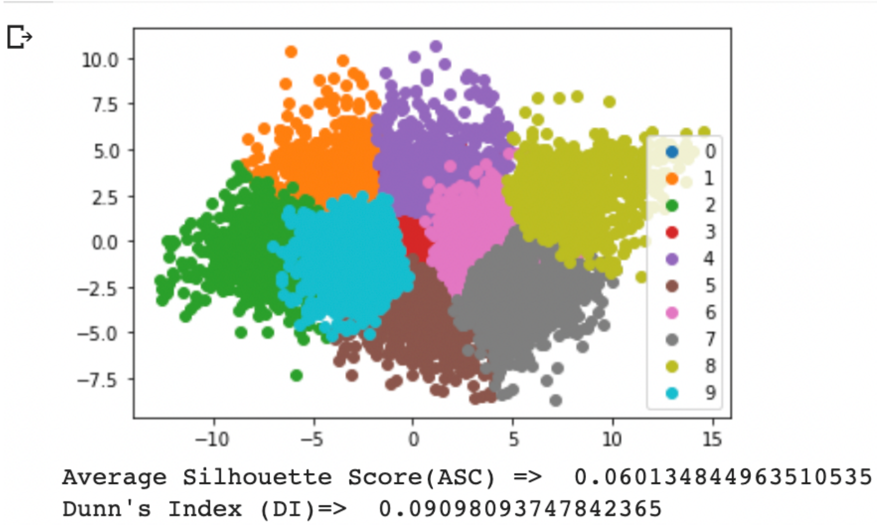
4 **K-Means Clustering** (**Training**)

K-means is based on the Euclidean distance and following is the algorithm:

- Define the initial 10 clusters, and their centers, a = a1, a2...a10 (randomly).
- As for every sample element, calculate the distance between element and every cluster center, and classify element into the cluster possessing the shortest distance.
- As for every cluster, re-calculate the cluster center,
- Repeat step 2 & 3 to reach a terminate condition.
- Using testing data set, quality of the clusters using ASC (Average Silhouette Coefficient) and DI (Dunn's Index) evaluation metrics.

5 Results for K-means:

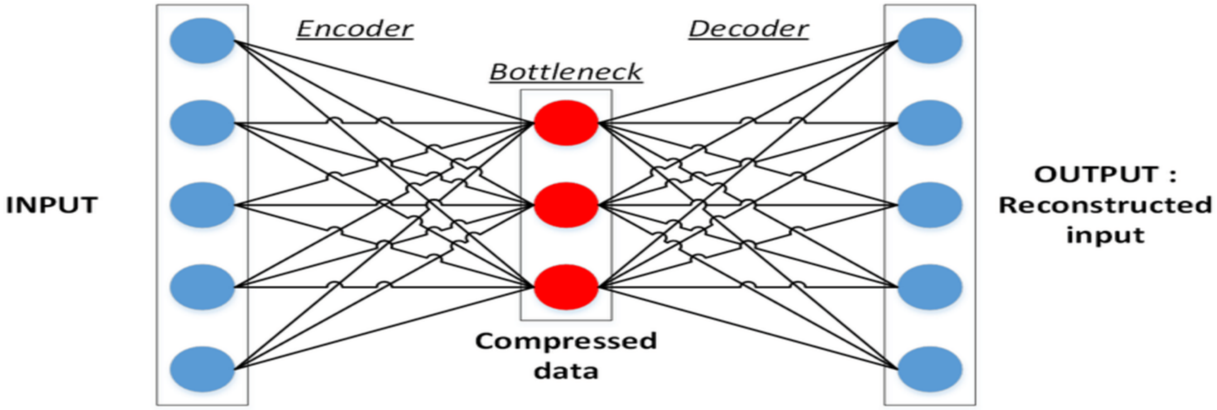
Above Trained model have been tested on test set of cifar 10 data set and plotted all examples (10,000) in below diagram and Evaluation has been done by Average Silhouette Score with 0.060 and Dunn’s Index value 0.090.



```
Average Silhouette Score(ASC) =>  0.060134844963510535
Dunn's Index (DI)=>  0.09098093747842365
```

6. **Auto -Encoders:**

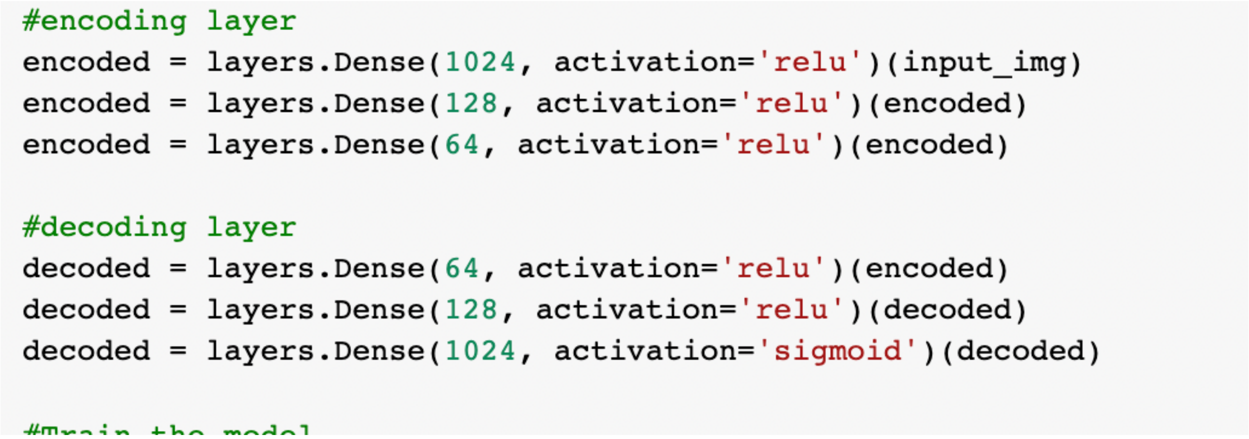
In general, an autoencoder consists of an encoder that maps the input x to a lower-dimensional feature vector z, and a decoder that reconstructs the input x^ from z. We train the model by comparing x to x^ and optimizing the parameters to increase the similarity between x and x^.



Training the Model:

Encoder: Learns how to compress the original input into a small encoding
Decoder: Learns how to restore the original data from that encoding generated by the Encoder.

The compressed code representations and adding optimizer's also aims to minimize the loss while reconstructing.



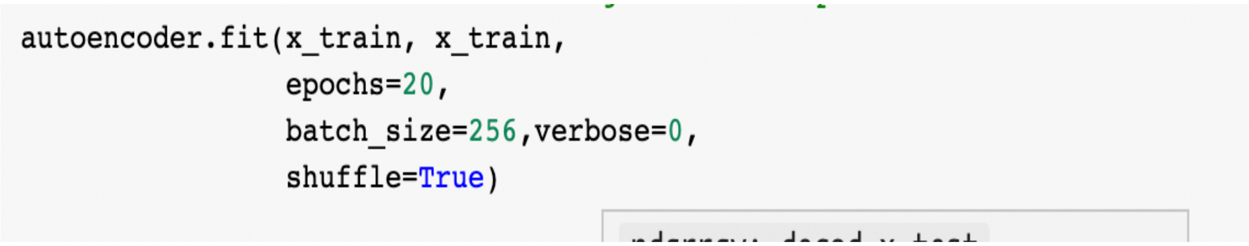
```
#encoding layer
encoded = layers.Dense(1024, activation='relu')(input_img)
encoded = layers.Dense(128, activation='relu')(encoded)
encoded = layers.Dense(64, activation='relu')(encoded)

#decoding layer
decoded = layers.Dense(64, activation='relu')(encoded)
decoded = layers.Dense(128, activation='relu')(decoded)
decoded = layers.Dense(1024, activation='sigmoid')(decoded)

#Train the model
```

Fit the Model:

After adding layers of encoders and decoders, then fit with training set of cifar10 dataset with number of epochs =20 and batch_size = 64 to get correct reconstructed input.
Then Predict the output by passing x_test parameters and generate the clusters of the elements using K-means defined from sparse representations generated from Auto-Encoders.



```
autoencoder.fit(x_train, x_train,
                epochs=20,
                batch_size=256,verbose=0,
                shuffle=True)
```

Results for Auto-Encoders:

Above Trained model have been tested on test set of cifar 10 data set and Evaluation has been done by Average Silhouette Score with range 0.06 – 0.08
Graph has been plotted on original images of cifar10 with reconstructed images after autoencoding.