Homework 3

Collaborators:
Name: Jessie Peng
Student ID: xxxxxxxxxx

Problem 3-1. Neural Networks

In this problem, we will implement the feedforward and backpropagation process of the neural networks.

(a) Answer: The testing loss is 0.247, and the testing accuracy is 0.922.

Problem 3-2. K-Nearest Neighbor

In this problem, we will play with K-Nearest Neighbor (KNN) algorithm and try it on real-world data. Implement KNN algorithm (in *knn.m*/*knn.py*), then answer the following questions.

(a) Try KNN with different K and plot the decision boundary.

Answer:

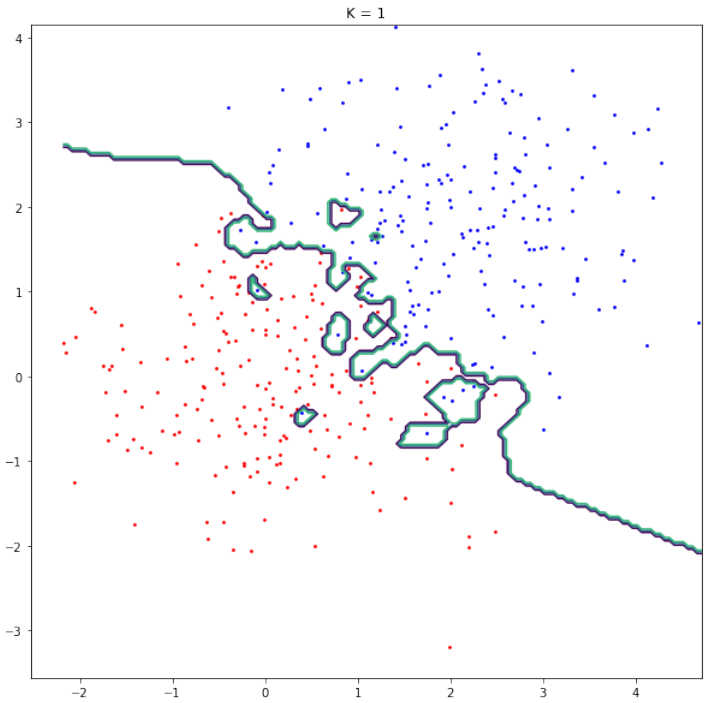

Figure 1: Decision boundary with k = 1

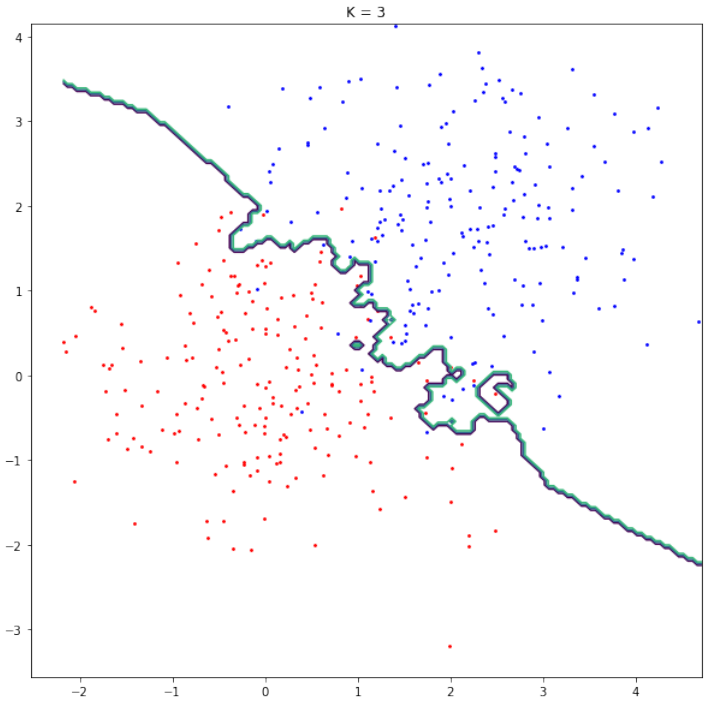

Figure 2: Decision boundary with k = 3

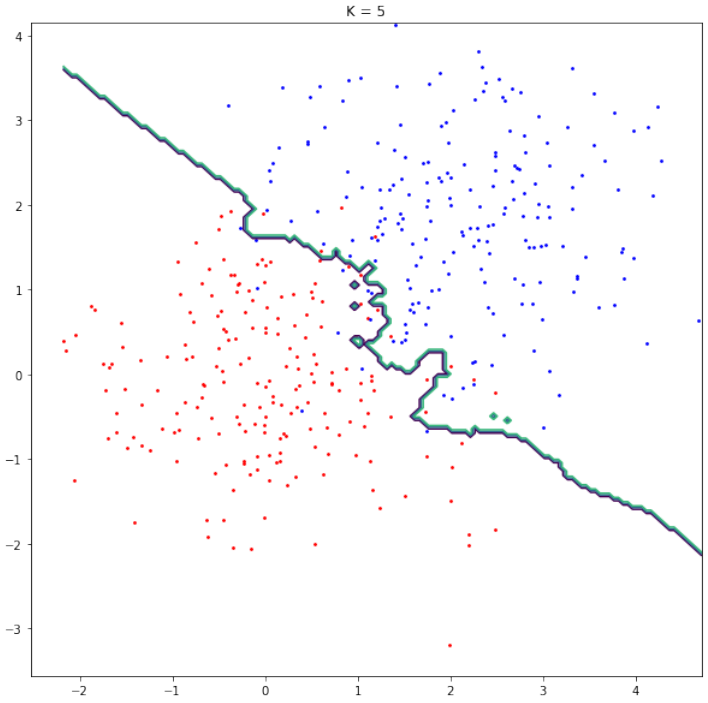

Figure 3: Decision boundary with k = 5

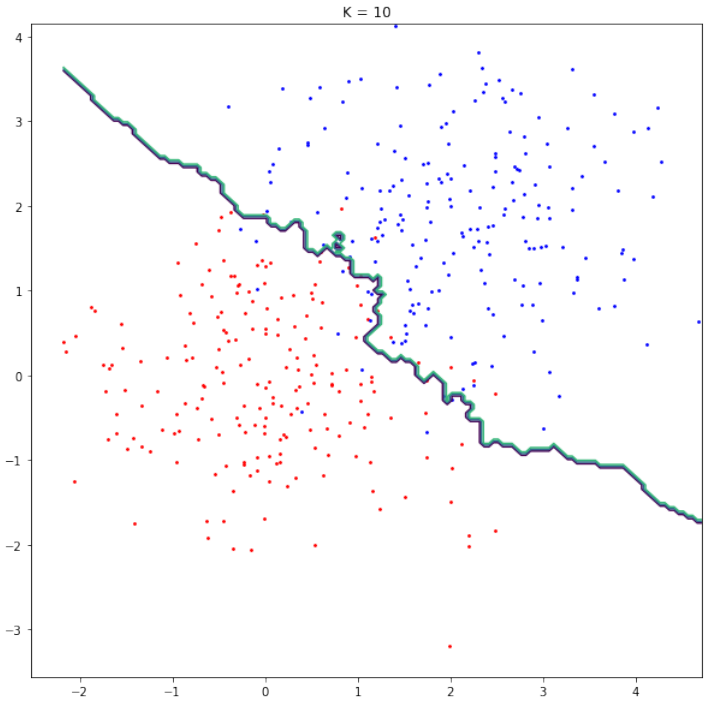

Figure 4: Decision boundary with k = 10

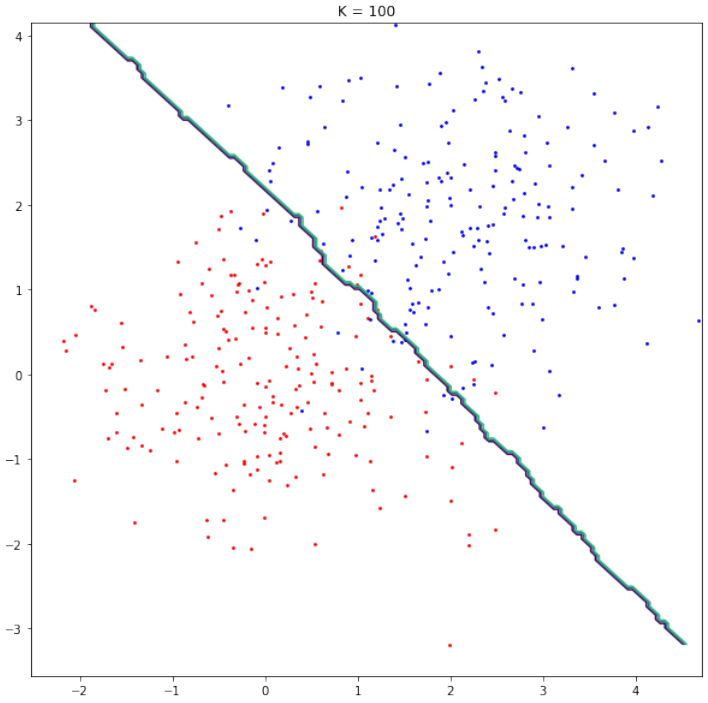

Figure 5: Decision boundary with k = 100

(b) We have seen the effects of different choices of K. How can you choose a proper K when dealing with real-world data ?

Answer: We can use techniques such as cross-validation to tune the parameter K - every time we pick a proper validation set from the training data and use different value of K to test them, then at last we choose the K that has the best overall performance during these validation process.

(c) Finish *hack.m*/*hack.py* to recognize the CAPTCHA image using KNN algorithm.

Answer: The training set contain images of 100 digits, which come from 20 CAPTCHA images processed by *extract_ image* helper function.

The test succeeded with the following test image:

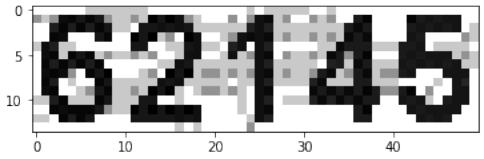

Figure 6: Test image ”62145”

Problem 3-3. Decision Tree and ID3

Consider the scholarship evaluation problem: selecting scholarship recipients based on gender and GPA. Given the following training data:

Answer: It’s quite obvious that GPA rather than Gender is the better attribute to pick first.

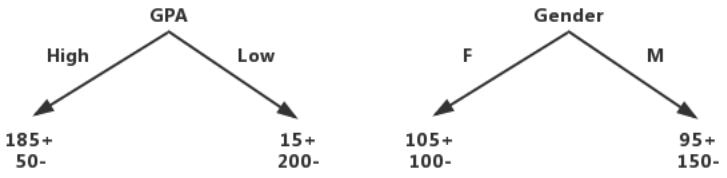

Figure 7: GPA is better than Gender as the first attribute

To calculate the information gain, we first calculate the entropy:

$$E_S = -\frac{200}{450}\log\frac{200}{450} - \frac{250}{450}\log\frac{250}{450} = 0.9911$$

$$E_H = -\frac{185}{235}\log\frac{185}{235} - \frac{50}{235}\log\frac{50}{235} = 0.7467$$

$$E_L = -\frac{15}{215}\log\frac{15}{215} - \frac{200}{215}\log\frac{200}{215} = 0.3651$$

$$E_{HF} = -\frac{95}{115}\log\frac{95}{115} - \frac{20}{115}\log\frac{20}{115} = 0.6666$$

$$E_{HM} = -\frac{90}{120}\log\frac{90}{120} - \frac{30}{120}\log\frac{30}{120} = 0.8113$$

$$E_{LF} = -\frac{10}{90}\log\frac{10}{90} - \frac{80}{90}\log\frac{80}{90} = 0.5033$$

$$E_{LM} = -\frac{5}{125}\log\frac{5}{125} - \frac{120}{125}\log\frac{120}{125} = 0.2423$$

Then we calculate the information gain of each internal node:

$$Gain(S, GPA) = E_S - \frac{235}{450}E_H - \frac{215}{450}E_L = 0.4267$$

$$Gain(S, High - Gender) = E_H - \frac{115}{235}E_{HF} - \frac{120}{235}E_{HM} = 0.0062$$

$$Gain(S, Low - Gender) = E_L - \frac{90}{215}E_{LF} - \frac{125}{215}E_{LM} = 0.0135$$

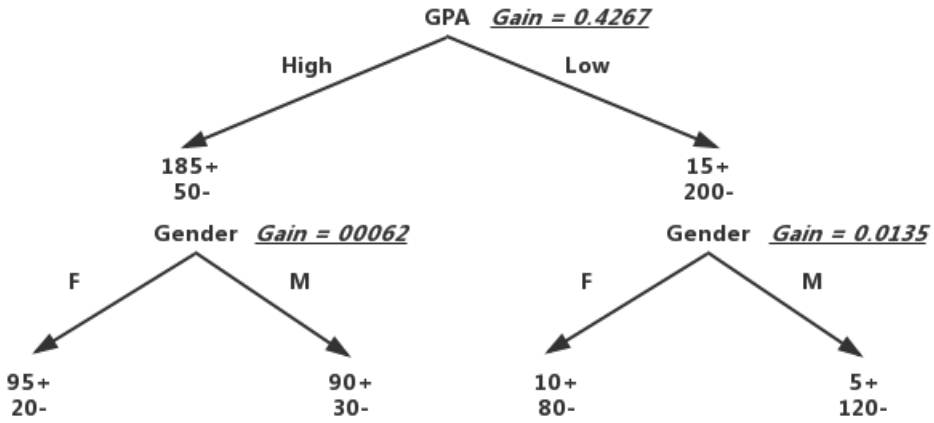

Figure 8: Decision tree with information gain

Problem 3-4. K-Means Clustering

Finally, we will run our first unsupervised algorithm k-means clustering.

(a) Visualize the process of k-means algorithm for the two trials.

Answer:

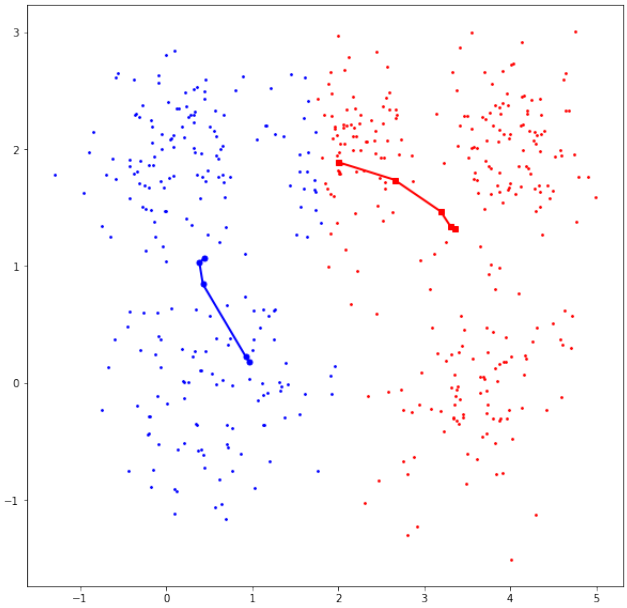

Figure 9: Clustering result with largest SD

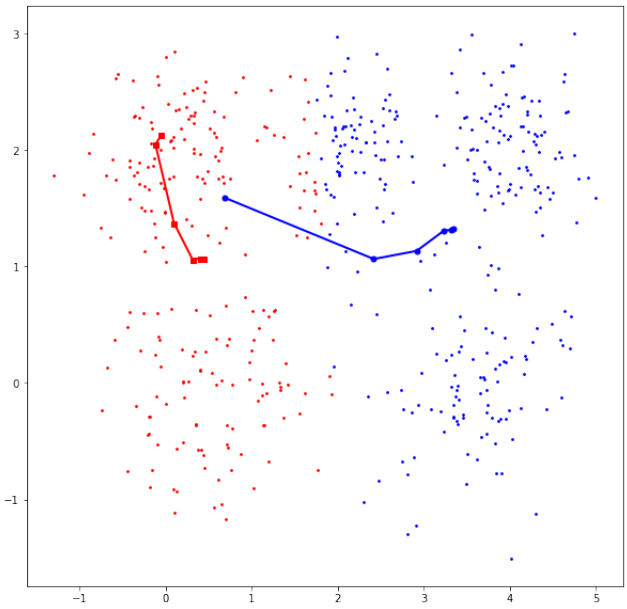

Figure 10: Clustering result with smallest SD

(b) How can we get a stable result using k-means?

Answer: We should run the k-means algorithm multiple times, and choose the clustering result that has the minimum sum of distances from each point to its respective centroid.

(c) Visualize the centroids.

Answer:

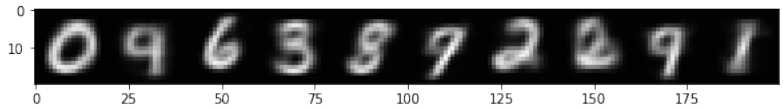

Figure 11: 10 centroids

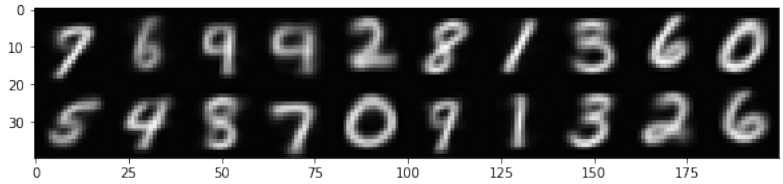

Figure 12: 20 centroids

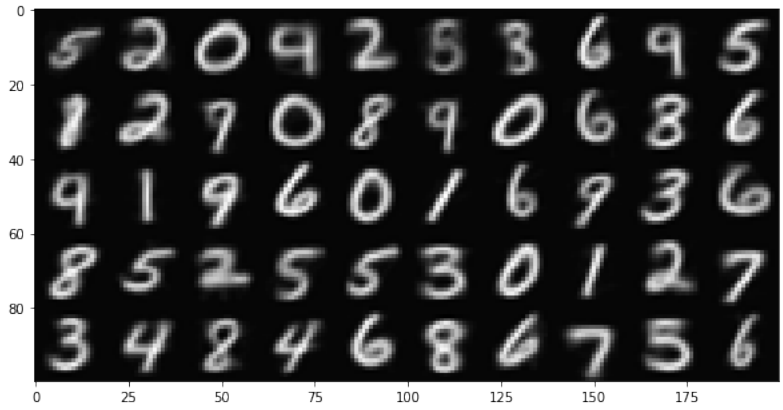

Figure 13: 50 centroids

(d) Vector quantization.

Answer:

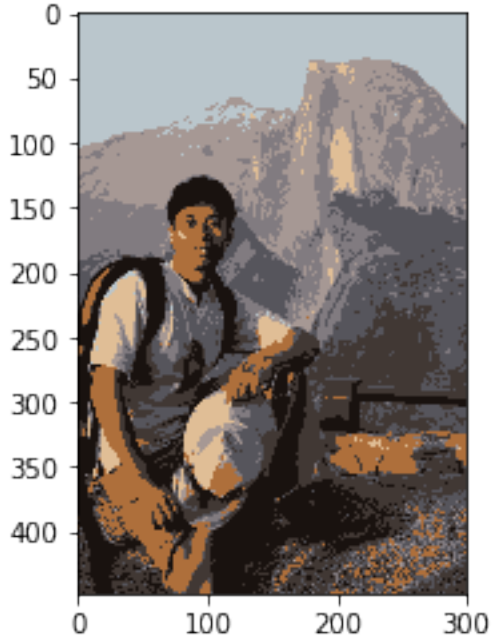

Figure 14: Compress Sample 0 with K = 8

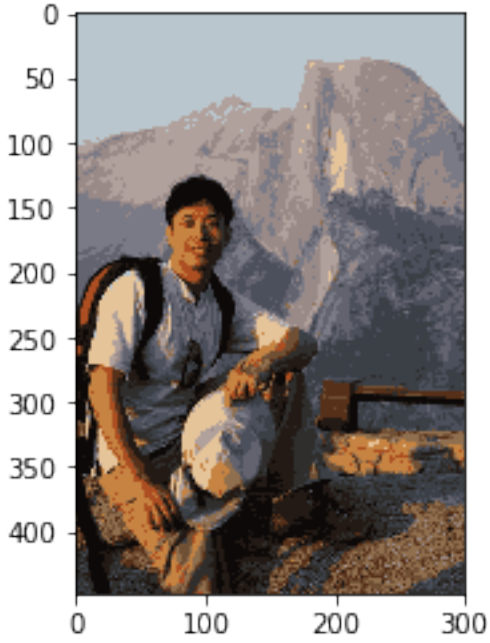

Figure 15: Compress Sample 0 with K = 16

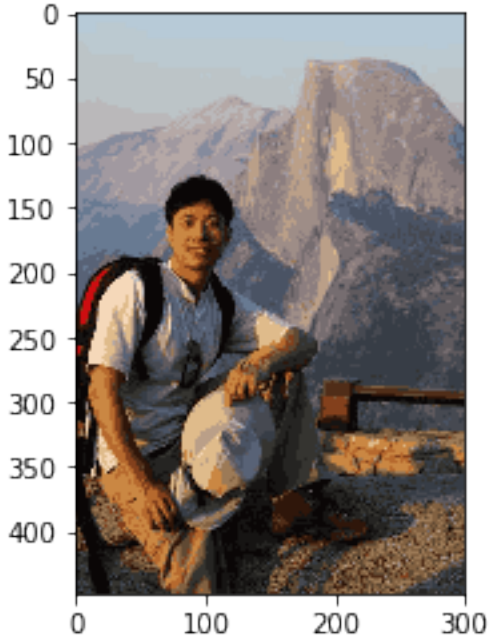

Figure 16: Compress Sample 0 with K = 32

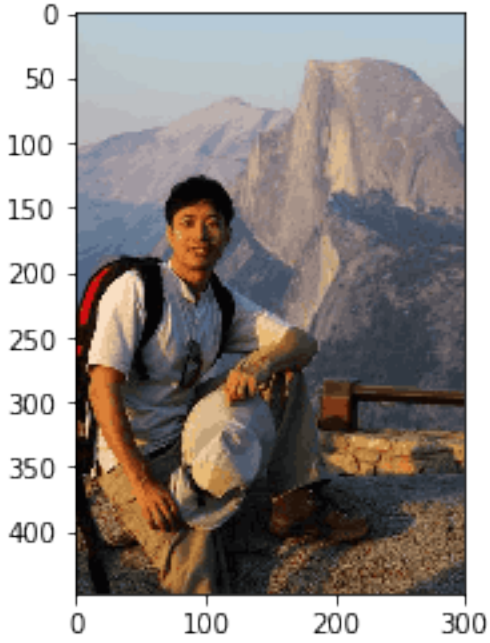

Figure 17: Compress Sample 0 with K = 64

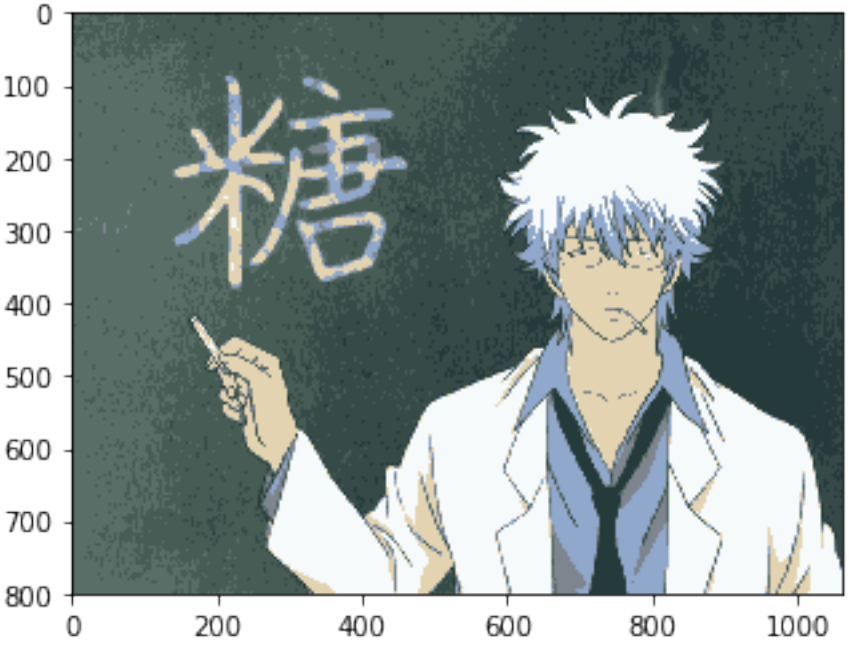

Figure 18: Compress Sample 1 with K = 8

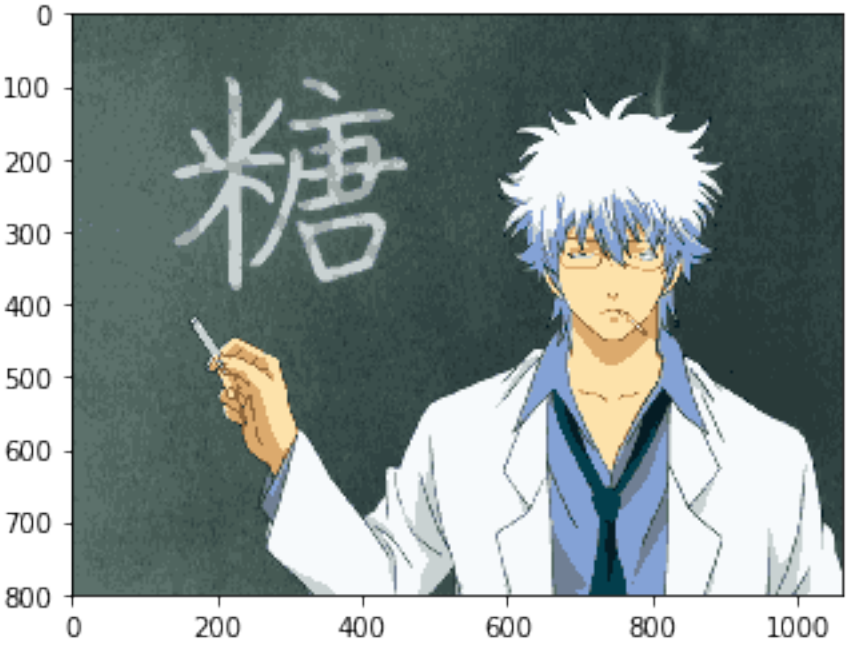

Figure 19: Compress Sample 1 with K = 16

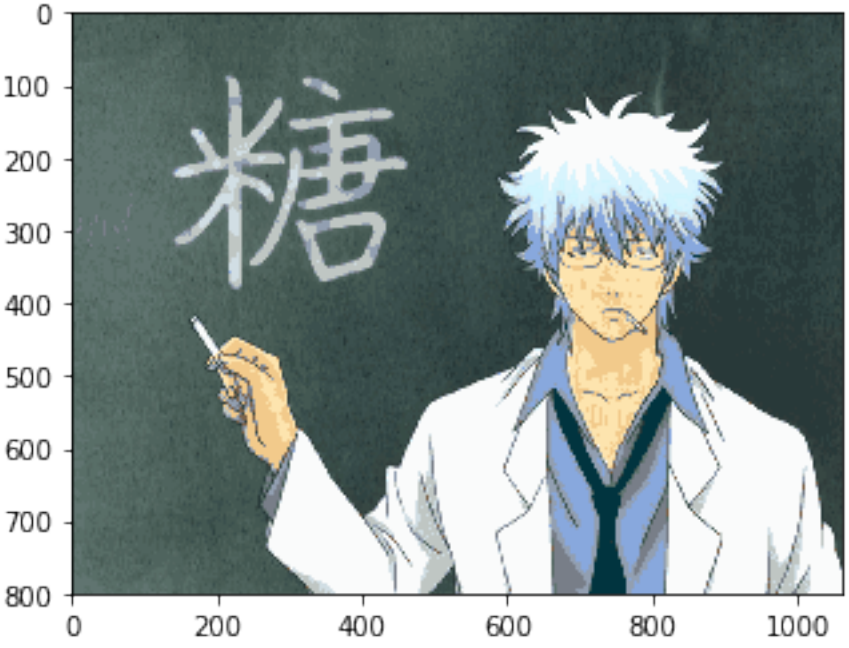

Figure 20: Compress Sample 1 with K = 32

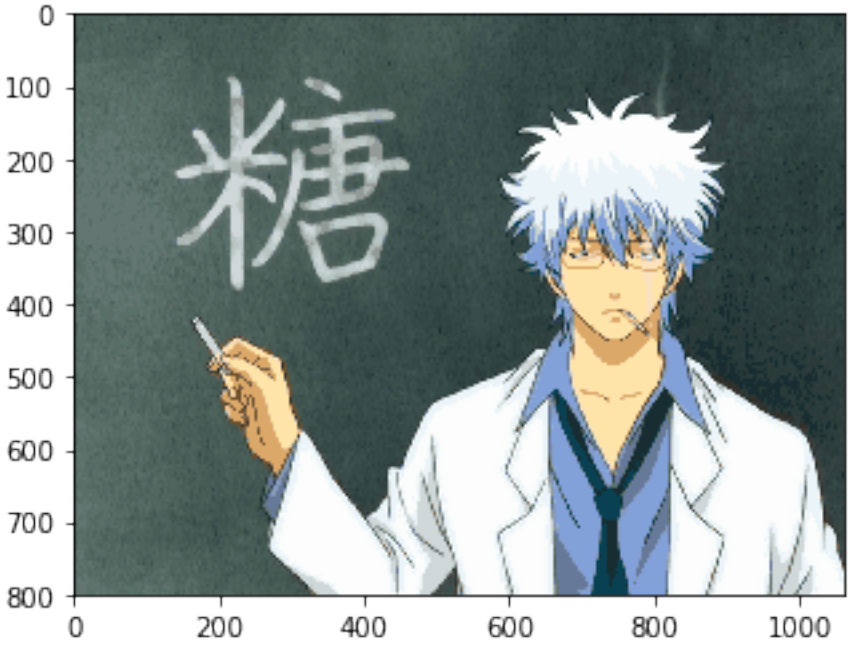

Figure 21: Compress Sample 1 with K = 64

When K = 64,

$$compress\ ratio = \frac{original\ data\ size}{compressed\ data\ size} = \frac{24}{\log 64} = 4.0$$