Homework 2 - Image Classification

Naomi Derel, 325324994 Sagi Ben Lulu, 207031493

18.12.2024

1 Classic Classifier

1.1 Loading Dataset

The first 5 images in the CIFAR-10 dataset are shown in Figure 1.

Figure 1: First 5 images in CIFAR-10 dataset

1.2 K-Nearest Neighbors

We load 10,000 samples from the training dataset, and concatenate the RGB channels of a flat image into a single vector. We then train a K-Nearest Neighbors classifier from `sklearn.neighbors` with $k = 10$.

1.3 Evaluation on Test Set

We load 1,000 samples from the test dataset, and evaluate the classifier on the test set. The accuracy of the classifier is 0.288. We also compute a confusion matrix between the true labels and the predicted labels, shown in Figure 2. We note that the results are not very good, with low accuracy and many misclassifications. The best performance was achieved on the 'ship' class, which might have to do with the distinguished background color or object itself.

1.4 Analysis of K Value

The comparison of the model accuracy as a function of the number of neighbors k, between 1 and 30, is shown in Figure 3. We note that the accuracy is pretty low for all values of k and remains below 0.3.

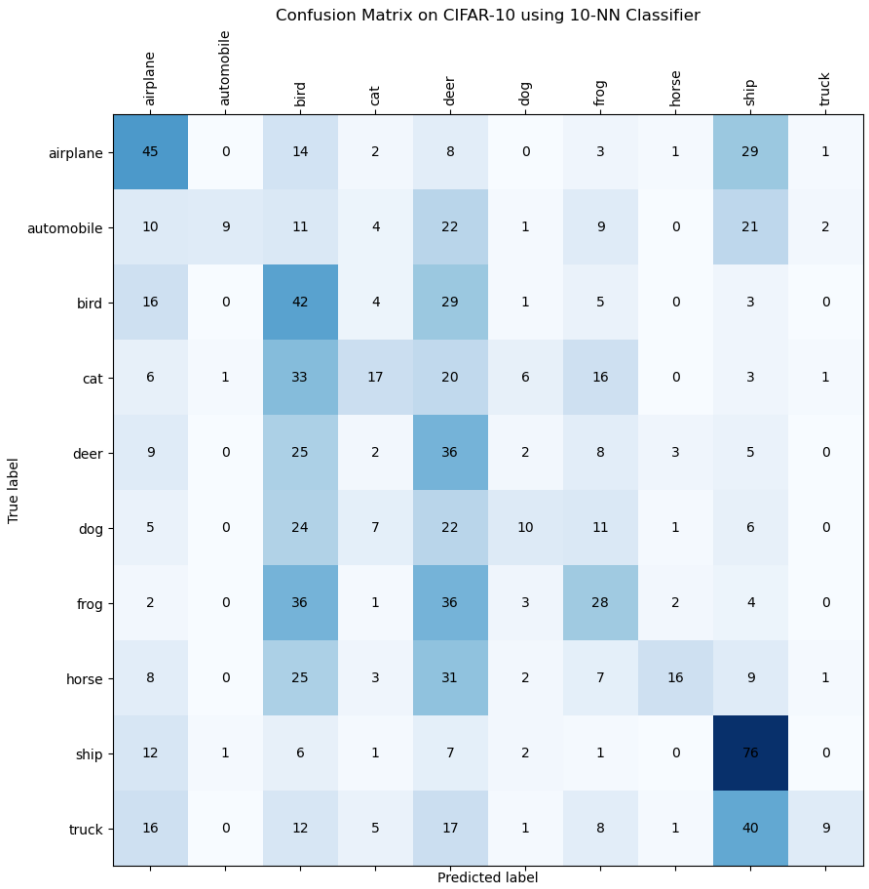

Figure 2: Confusion matrix of K-Nearest Neighbors classifier

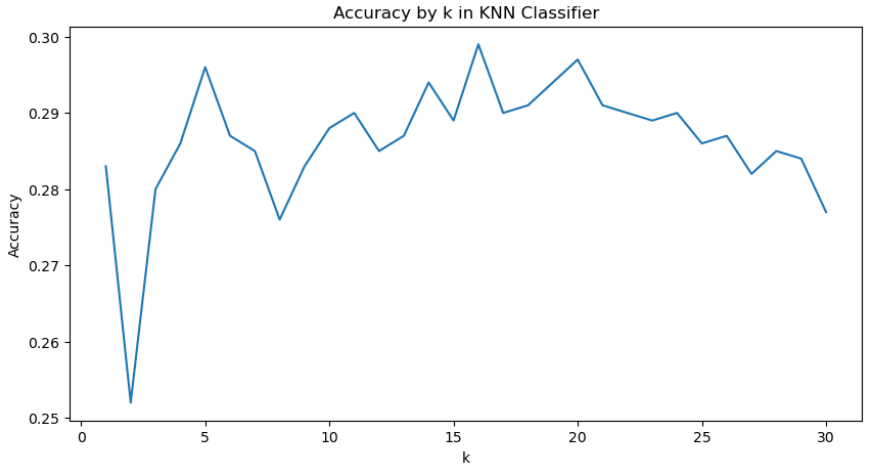

Figure 3: Model accuracy as a function of the number of neighbors k

2 Question 3 - Foundation Models

2.1 Embedding Space of T-SNE

We load all the images from the 'data/clip_images' directory, which includes 'cats' and 'dogs' images. We then embed the images using clip, and project them using T-SNE (with random seed=18). The resulting plot is shown in Figure 4. We note that the 'cats' and 'dogs' images are well separated in the embedding space, with the 'cats' images on the right and the 'dogs' images on the left.

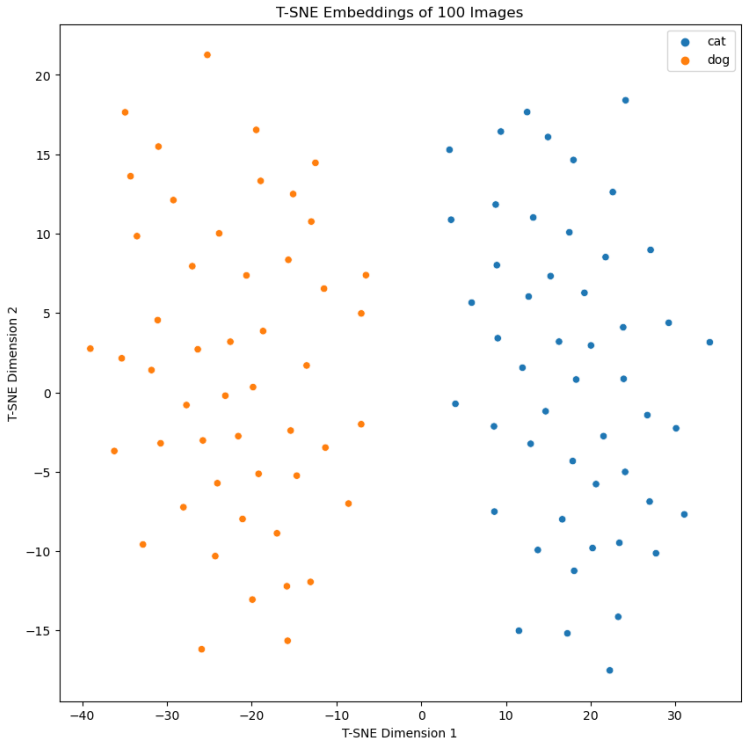

Figure 4: T-SNE plot of the 'cats' and 'dogs' images

2.2 CLIP Nearest Neighbors

We compare the image of Alfie (Figure 5) by cosine similarity in the embeddign space (Figure 6).

Figure 5: Alfie

Figure 6: 5-Nearest Neighbors of Alfie in the CLIP embedding space

First, the results are all cat images and no dog images, which matches our expected result. Second, we notice that most of the cats similar coloring to Alfie (gray overall or white in the same spaces). This is a good indication that the model is able to understand the context of the image and not just the image itself.

2.3 Classification with CLIP Textual Embeddings

We load the images and labels from the 'data/clip_images/test' directory. WE embed all the images using CLIP, and create the labels: "A photo of a dog", "A photo of a cat". We embed the labels as well using the CLIP textual model. We then compute the cosine similarity between the image and the labels, and classify the image based on the label with the highest similarity.

We achieve an accuracy of 100% on the test set, and the following confusion matrix:

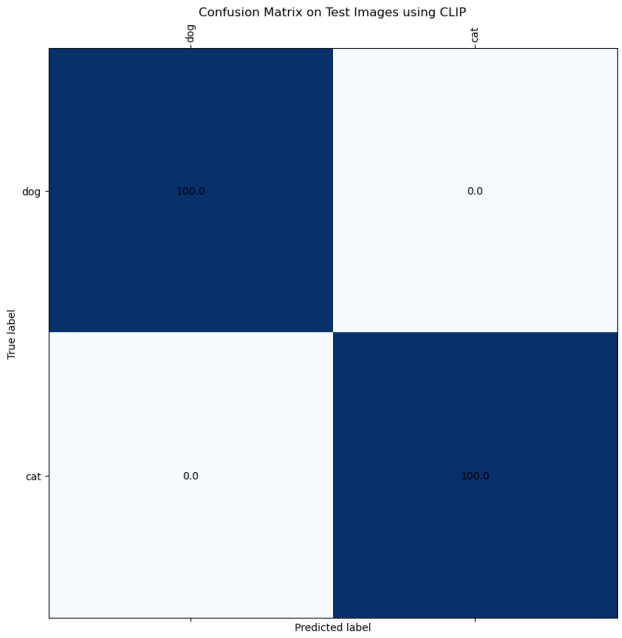

Figure 7: Confusion matrix of the classification with CLIP textual embeddings

2.4 Counting Objects in Images

We write a function similar to the last step, and define the labels: "A photo with zero cats", "A photo of a cat", "A photo of 2 cats",etc, up to 10 cats, which should cover reasonable images as we saw in the dataset. We then classify the images based on the label with the highest similarity.

We succeed on 2 simple cases, with 1 and 2 cats (Figures 8, 9). However, on an image with a cat and a human (10), we also get the prediction of 2 cats, as well as an image of a dog where we get a prediction of 1 cat. This indicates that the model might not grasp the label of zero cats as intended.

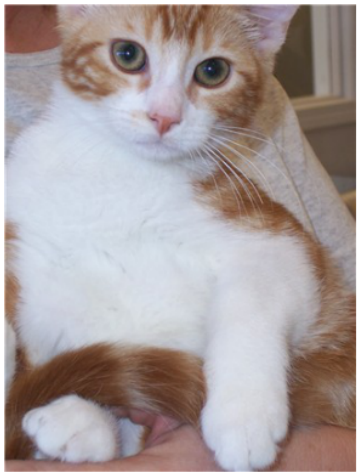

Figure 8: Predicted 1 cat

Figure 9: Predicted 2 cats

Figure 10: Predicted 2 cats