CLIP-based Image Classifier

April 18, 2025

Abstract

In this project, we will implement a classifier based on the CLIP model. The model receives an image with a description and predicts its corresponding class. We compare the model's zero-shot and linear probe performance in classifying the CIFAR-10 dataset.

1 Implementation

For the CLIP model, I used the Open CLIP ViT-B-32 implementation, which is available at Open CLIP repository.

For the linear probe, I used a logistic regression classifier, and for zero-shot, I used the cosine similarity between the extracted features from the input image and the extracted features from class names. The code with complete comments is available in the Google Colab notebook.

2 CIFAR-10 dataset

This dataset includes 10 classes, including 50000 images in the training set and 10000 in the test set.

3 Evaluation

3.1 Compare linear probe and zero-shot performance

For comparison, we evaluated each method on the CIFAR-10 test dataset and made a comparison based on the accuracy metric.
As shown in Table 1, the linear probe achieves an accuracy that is 3.08% higher than that of zero-shot CLIP. This is expected, as the linear probe is trained in a fully supervised manner using the dataset labels, whereas the zero-shot performance relies solely on CLIP's pre-trained features without any additional training.

Method	**Accuracy**
Zero-Shot	93.66%
Logistic Regression	96.74%

Table 1: Accuracy comparison of CLIP-based classification methods on CIFAR-10

3.2 Explore class names influence on the model's performance

To explore the influence of class names on the model's performance, I used three different text input methods for each class:

- **First method**: Simply using the class names as a single word, e.g. "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", "truck".

- **Second method**: Using the pattern ”a photo of a [class name]”, resulting in inputs like ”a photo of an airplane”, ”a photo of an automobile”, ”a photo of a bird”, ”a photo of a cat”, ”a photo of a deer”, ”a photo of a dog”, ”a photo of a frog”, ”a photo of a horse”, ”a photo of a ship”, and ”a photo of a truck”.

- **Third method**: Using more descriptive phrases, such as ”a realistic picture of an aircraft”, ”a modern vehicle on the road”, ”an image of a flying bird”, ”an image of a pet cat”, ”an image of a deer standing in grass”, ”a realistic photo of a dog”, ”a close-up of a small frog”, ”a wild horse in nature”, ”an image of a ship at sea”, and ”a realistic picture of a truck”.

Method	**Class names**	**Accuracy**
Zero-Shot	first method	71.75%
Logistic Regression	first method	96.74%
Zero-Shot	second method	93.66%
Logistic Regression	second method	96.74%
Zero-Shot	third method	87.70%
Logistic Regression	third method	96.74%

Table 2: Influence of the class names on the accuracy of the model

It can be seen in Table 2 that the linear probe still outperforms the zero-shot approach due to its fully-supervised training with the dataset labels. However, the performance gap increases as the text prompts become less detailed. For example, in the first method, we used just one word to define the class name. In the second method, we used the pattern ’a photo of an object’, which is closer to the pattern the CLIP model was trained on. In the third pattern, the class names remain close to the CLIP training dataset, but unnecessary words like ’modern’ in ’a modern vehicle on the road’ are introduced. Based on this, the second method demonstrates the best performance for both the linear probe and zero-shot approaches.