Project Documentation

Image classification: Fields and Roads

Author:

Badr Bouzinab

July 2023

Contents

1 Introduction:

In the field of image classification, Convolutional Neural Networks (CNN) have been the classical method for achieving remarkable results. Recently, transfer learning has emerged as a powerful technique that enables us to transfer the knowledge learned from one task and apply it to another related task.

In this project, we will take advantage of the power of transfer learning and pre-trained models. We will use a pre-trained model CLIP (pre-trained on 400M pairs of image-text) to extract images embeddings and use them as input to a simple Multi-Layer Perceptron (MLP) for binary classification.

2 Model Architecture:

CLIP (Contrastive Language-Image Pre-Training) is a multi-modal vision and language model. it uses an image encoder and a text encoder to generate respective embeddings of identical dimensions. Then, the image-text similarity is measured through the dot product operation between the image and text embeddings.

In our project, we take advantage of CLIP's image encoder to extract image embeddings, which are then fed into an MLP consisting of two linear layers and a ReLU activation function to predict image class.

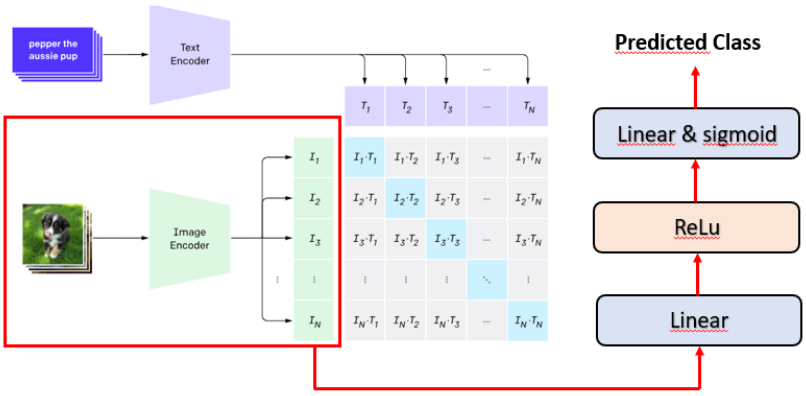

Figure 1: Model architecture overview.

The initial linear layer transforms the input embedding obtained from CLIP into a 256-dimensional feature representation, followed by the application of the Rectified Linear Unit (ReLU) activation function. Then, the second linear layer further processes the 256-dimensional vector to produce a scalar output. The final step involves passing the scalar through the sigmoid function, leading to a binary decision: if the sigmoid output is greater than or equal to 0.5, the predicted class is 1; otherwise, it is 0.

3 Training result:

During training, we employ binary cross entropy loss. We use adam optimize with an initial value of learning rate equals 3.10^{-3}. We set the batch size to 3.

We stop training after 7 epochs, where the training accuray reaches 98.7%

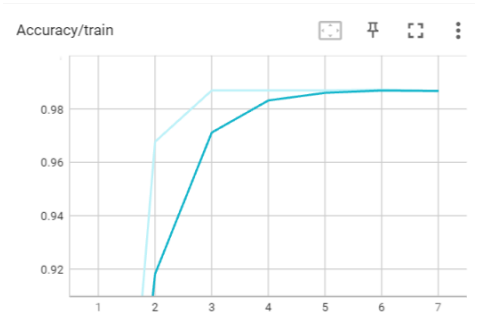

Figure 2: Training accuracy curve, using tensorboard.

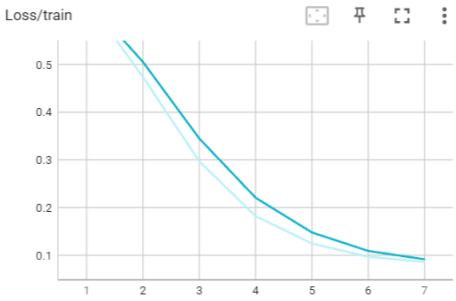

Figure 3: Training loss curve, using tensorboard.

4 Comparaison between our model and CLIP model:

In this section, we conduct a comparative analysis of model predictions using the test set images. We evaluate our model and the CLIP model. By employing 2 captions: "this is a photo of a field" and "this is a photo of a road.", CLIP calculates the similarity score between the image and each caption, and the predicted class is determined based on the caption with the highest score.

On the 10 images from the test set, both models yield identical predicted classes in 8 instances. However, in the remaining 2 images, the two models output different predictions:

Figure 4: image 1, CLIP prediction: field, our model prediction: road.

Figure 5: image 6, CLIP prediction: field, our model prediction: road.

Indeed, the images present inherent ambiguity even to human observers. The first image presents a trampled path in the middle of a field. raising the question of whether it qualifies as a road. The second image presents both a road and field. should we take the angle of taking the photo into consideration in the classification process.

To address this ambiguity, we may consider two potential solutions:

- Introduce a Third Class: By incorporating a third class, labeled as "road and field," we transform the problem into a multi-class classification. This approach enables the model to distinguish between images portraying a clear road and field, and images containing elements of both.
- Probabilistic Analysis: we can evaluate the probabilities of belonging to each class. For instance, the model assigns a score of 0.54 to image 1 and 0.6 to image 6, both of which are close to the threshold of 0.5. Consequently, we can segregate images with scores around 0.5 into a separate category for further examination.

5 Potential Approaches for Enhancing Model Generalization:

As the training dataset is relatively small, to improve the generalization of the model, we can use the following startegies:

- Apply data augmentation to increase the dataset size. In our case, this augmentation process can addresse the dataset imbalance issue (45 field images and 108 road images). Specifically, augmentation can be applied more extensively to the minority class, in order to ensure a balanced representation of both classes.
- Increase the number of epochs, and apply learning rate decrease startegies: e.g we can apply a linear decay approach such that the learning rate value drop by half after a predefined number of epochs.
- To avoid overfitting, we can apply an early stopping strategy, where we save the best checkpoint after each validtion step. If the model did not imporve after a certain number of epochs, we stop the training.