Machine Learning - HA7

Weili Diao Student ID: 21127071

Program: Big Data Techonology Email: wdiaoaa@connect.ust.hk

1 Introduction

1.1 Project Objective

The task of this assignment is to perform zero-shot prediction on the chosen dataset using one CLIP model, while implementing zero-shot prediction with two different sets of prompt templates: simple template and ensemble template.

This report begins by introducing the dataset and CLIP model selected for the task, followed by an explanation of the methodology used, including the prompt templates and how multiple templates were applied each class. Then it compares the classification accuracy achieved with each set of templates and analyzes the findings. Finally, visualizations of several predictions are presented.

1.2 Summary

1.2.1 Model & Dataset

For this task, we use the ViT-B/32 version of the CLIP model, which is based on the Vision Transformer architecture.

Then, we apply the CLIP model to the RESISC45 dataset, which contains 45 classes, including categories like 'airplane', 'beach', 'forest', 'stadium', and many more, which represent different types of landscapes, urban areas, and natural features.

1.2.2 Environment

Kaggle: GPU T4 × 2
Python version: 3.10.14
Operating System: Linux 5.15.154+
PyTorch version: 2.4.0+cu123

2 Experiment

2.1 Methodology

The methodology for applying the CLIP model to the RESISC45 dataset involves two types of prompt templates: Simple Templates and Ensemble Templates. These templates are used to generate textual descriptions for each class, which are then encoded into text features by the CLIP model.

2.1.1 Simple Template

The Simple Template is a straightforward textual description for each class. Each class is associated with a single prompt in the format:

```
"a photo of a {class_name}"
```

Then CLIP uses this template to generate text embeddings. The simple templates for all 45 classes are created by iterating over the list of class names from the RESISC45 dataset. The generated simple templates are tokenized and passed through the CLIP model to obtain their text embeddings.

2.1.2 Ensemble Template

The Ensemble Template approach expands on the simple template by using multiple varied descriptions for each class. This method aims to provide a more comprehensive textual representation of each class. For each class, we generate a set of descriptive phrases. The templates include phrases:

```
"satellite imagery of {}.",
"aerial imagery of {}.",
"satellite photo of {}.",
"aerial photo of {}.",
"satellite view of {}.",
"aerial view of {}.",
"satellite imagery of a {}.",
"aerial imagery of a {}.",
"satellite photo of a {}.",
"aerial photo of a {}.",
"satellite view of a {}.",
"aerial view of a {}.",
"satellite imagery of the {}.",
"aerial imagery of the {}.",
"satellite photo of the {}.",
"aerial photo of the {}.",
"satellite view of the {}.",
"aerial view of the {}."
```

The ensemble approach creates a list of prompts for each class, and the corresponding textual features are generated by tokenizing and encoding each prompt. Each class is associated with a set of these templates, and for each template, the CLIP model generates a corresponding text feature. The main implementation strategy is to average each class: for the ensemble templates, the mean of the text features across all prompts for each class is calculated to create a single, representative feature for each class.

Generally, the simple templates provide a straightforward approach, while the ensemble templates aim to enhance classification accuracy by incorporating more varied and specific descriptions for each class.

2.2 Results

According to the experiment, we can get that the overall accuracy of Simple Template is 53.13% while the overall accuracy of Ensemble Template is 57.95%. This suggests that the use of multiple diverse prompts in the Ensemble Template provides some advantage in improving classification performance.

The following line chart (Figure 1) shows the prediction accuracy across batches during the experiment. In this experiment, we set the batch_size of the DataLoader to 700, which is the same as the number of images per class. Therefore, each batch contains images from a single class.

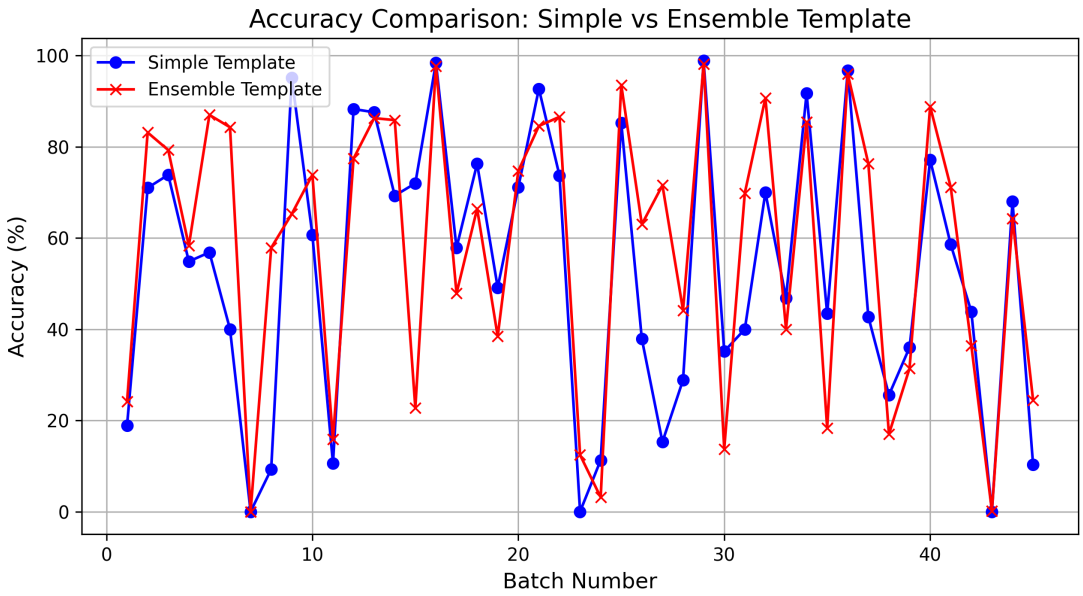

Figure 1: Accuracy Comparison

For both methods, there are some batches where accuracy is high, close to 100%, such as golf course and parking lot. However, there are also batches with very low accuracy, even approaching 0%, such as chaparral and meadow. This could indicate that the images in these classes are challenging to differentiate, possibly due to similar visual characteristics or ambiguous contexts.

Generally, for many of the batches, the Ensemble Template appears to perform better than the Simple Template. And the Simple Template is more likely to experience extreme accuracy fluctuations, whereas the Ensemble Template offers a smoother performance. However, in some classes, the Simple Template still outperforms the Ensemble Template, such as circular farmland.

2.3 Visualization

Figure 2 & 3 are some examples of misclassification from the two templates.

Figure 4 & 5 are some examples of circular farmland class from the two templates, in which the Simple Template performs much better than the Ensemble Template.

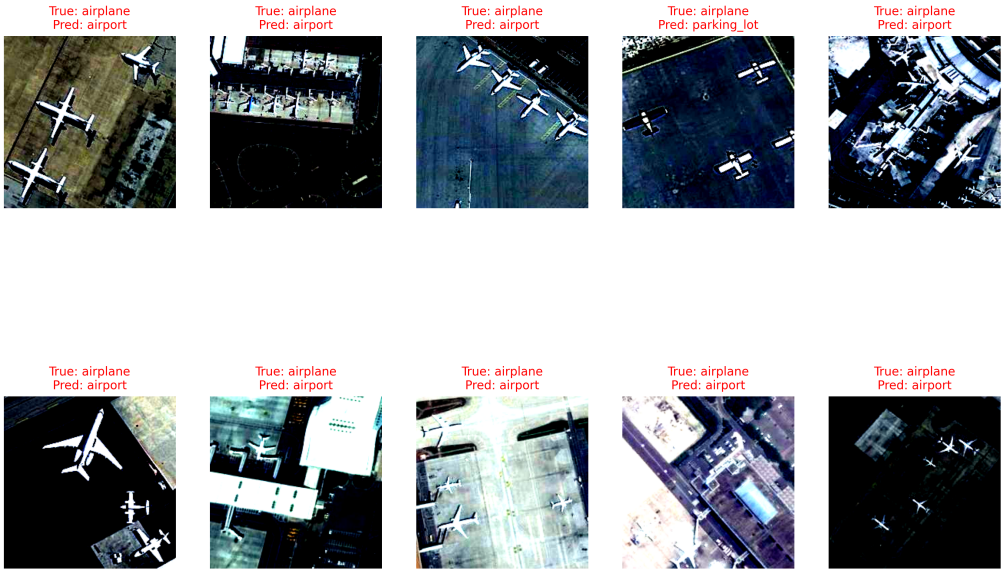

Figure 2

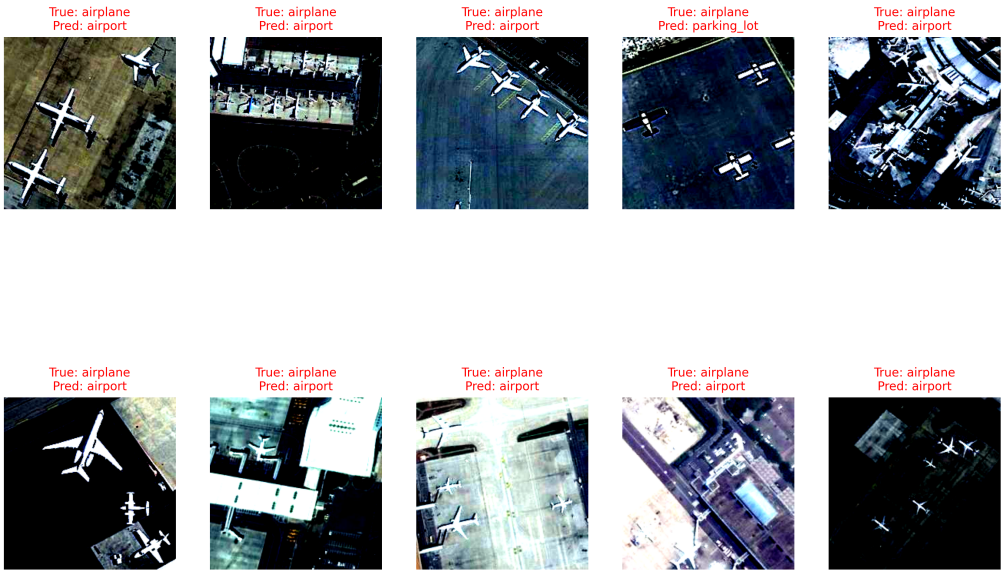

Figure 3

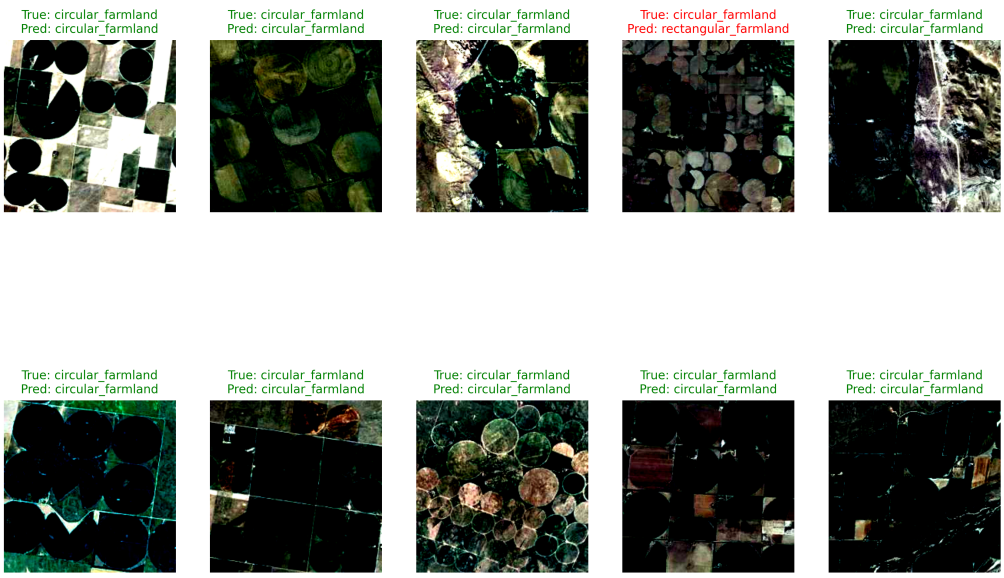

Figure 4

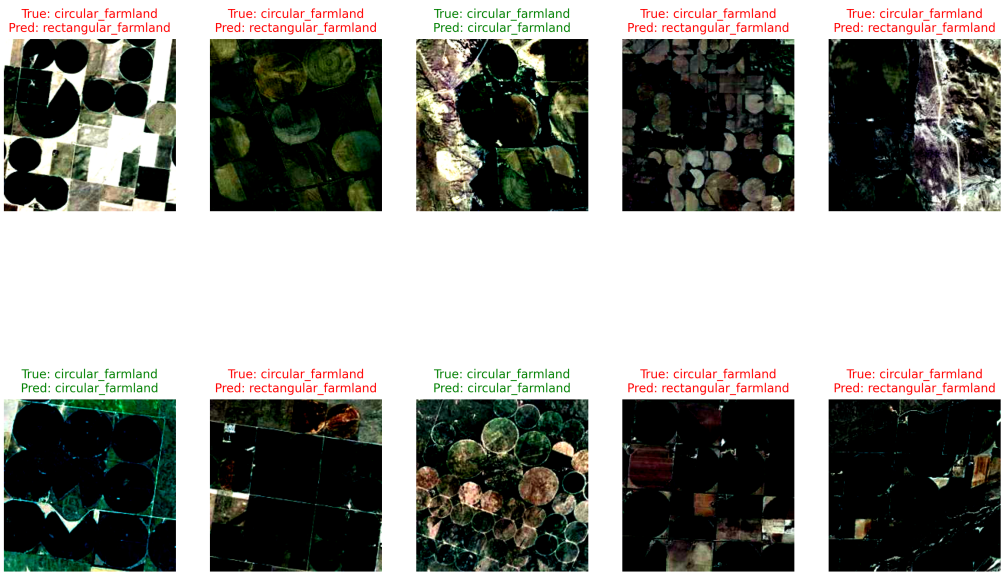

Figure 5