Trimble / Bilberry : AI Engineer technical exercise

Théo Cayla

Task: Create a two-class classifier: Field & Road

To address this task, the chosen strategy is to use either a CNN or a Vision Transformer, pretrained on the Imagenet classification task, as both have proven to be well performing to solve such problems.
Regarding the performance achieved, both models gave an accuracy of 90%, as one image is systematically mis-classified in both cases. In the case of the ViT, the misclassified image (*6.jpeg*) contains both a road and a field, which can explain the confusion of the model.

Dataset

The trainset is constituted of 111 images of road and 43 images of field. The quantity of training data is too small to train models as large as ViT or ResNet from scratch, that is why using the pretrained network is a necessity. Image augmentations are used to enlarge the diversity of training examples and help the model generalize.
The test set is constituted of 10 images, 5 of each class. This gives a rough estimation of the model accuracy, but isn't quite enough to properly evaluate its ability to generalize on unseen examples.
We observe a class imbalance in the trainset, with the road class being represented almost three times more than the field one. By upsampling the minority class, we'll see in the result section that we get a faster a more stable convergence of the model.

A custom class is created for the dataset, it allows to instantiate a dataset from the data directories. It inherits from the pytorch ImageFolder class, with a rewriting of the __getitem__ method, in order to be able to retrieve image names during testing.

Training Parameters

Several CNN architectures were tested for this task, Resnet-50 showed the best performances on this classification task, hence was selected. Tests were done using the Vision Transformer, they gave a similar accuracy than the ResNet, with a slightly faster convergence. Given the limited quantity of examples in the training set, the convolutional layers are all frozen during the fine tuning, and only the last fully connected layer is replaced and trained on the field/road images. Training parameters were determined empirically, finding a combination that offers the best trade-off regarding final accuracy, training speed, and accuracy stability when the training loss converges.
The parameters selected are:

- Adam optimizer
- Cross entropy loss
- Batch size: 50
- Learning rate: 0.001 (the use of a scheduler has proven to be ineffective in this scenario)
- Number of epochs : 100, even though the loss still decreases, the accuracy reaches 0.9 quite quickly after the training starts. Training for a longer time might result in

overfitting. Implementing an early stopping method would help prevent overfitting, but more test examples are necessary to detect and control that behavior.

Results

The trained model shows an accuracy of 0.9, meaning that it properly classifies 9 out of the 10 images of the test set. Given the few numbers of test examples we have, we can analyze those results in details. A method is implemented to count the number of mis-predictions for each image of the test set, during the whole training process. We can see that one image is constantly mispredicted, but it differs when using the CNN or the ViT. The CNN misclassifies field image *4.jpeg*, further investigation should be made to understand why this image fails the model, as its content is not out of distribution, and its format is similar to other images of the test set. The ViT misclassifies the image *6.jpeg* which both contain a road and a field, and could be a source of human misclassification for sure.

One lead to improve this performance is to used Grad-CAM technique to plot the activation map of the prediction, in order to get a sense of why those images are problematic in each case.
In the case of the ViT, one way to solve this issue would be to introduce similar examples in the trainset, containing both roads and fields. Plenty should be found in StreetView-like datasets.
Another idea is to approach the task differently, and consider it a multi label classification task, as it is possible that some images contain both a road and a field.

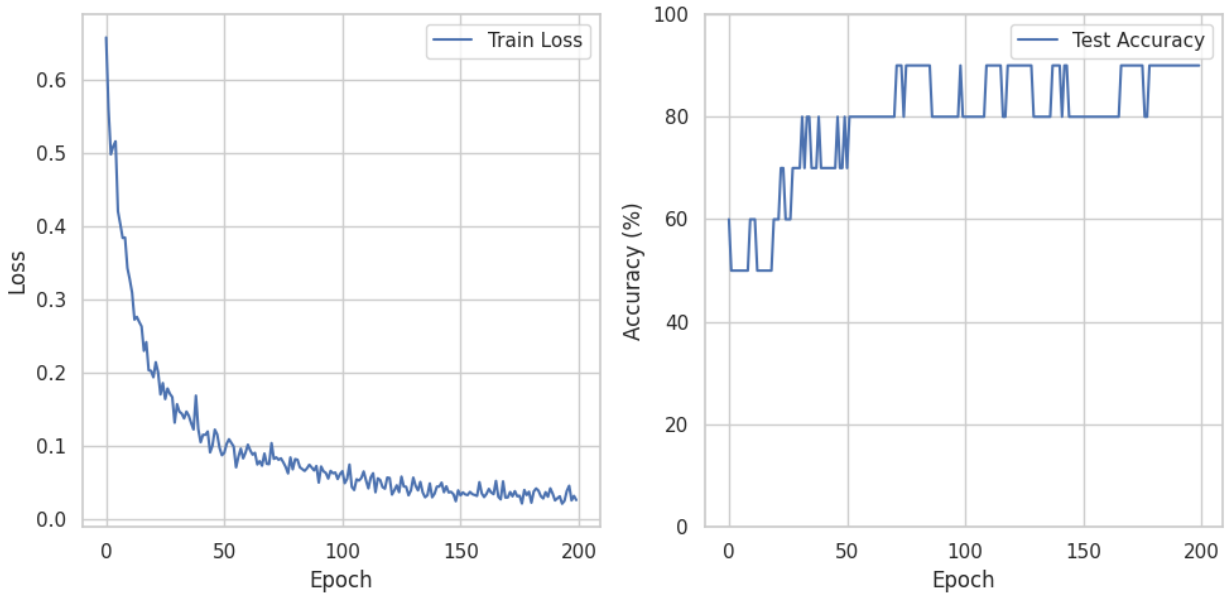

Figure 1 - Train loss and model accuracy of the model trained on the initial dataset

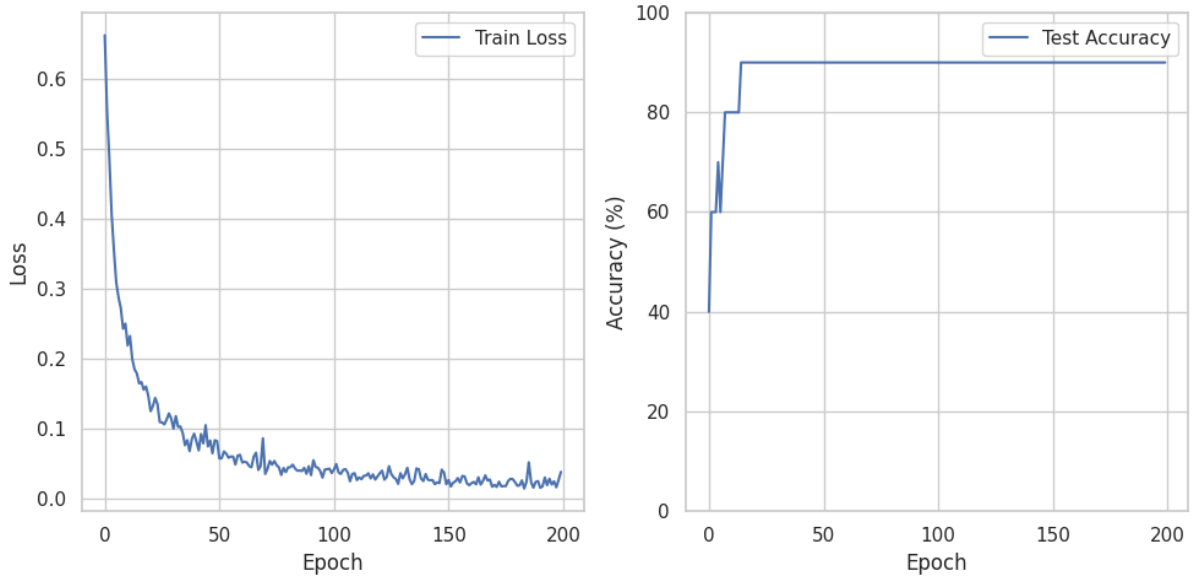

Figure 2 - Effect of oversampling the minority class, better accuracy is reached quicker and stays stable

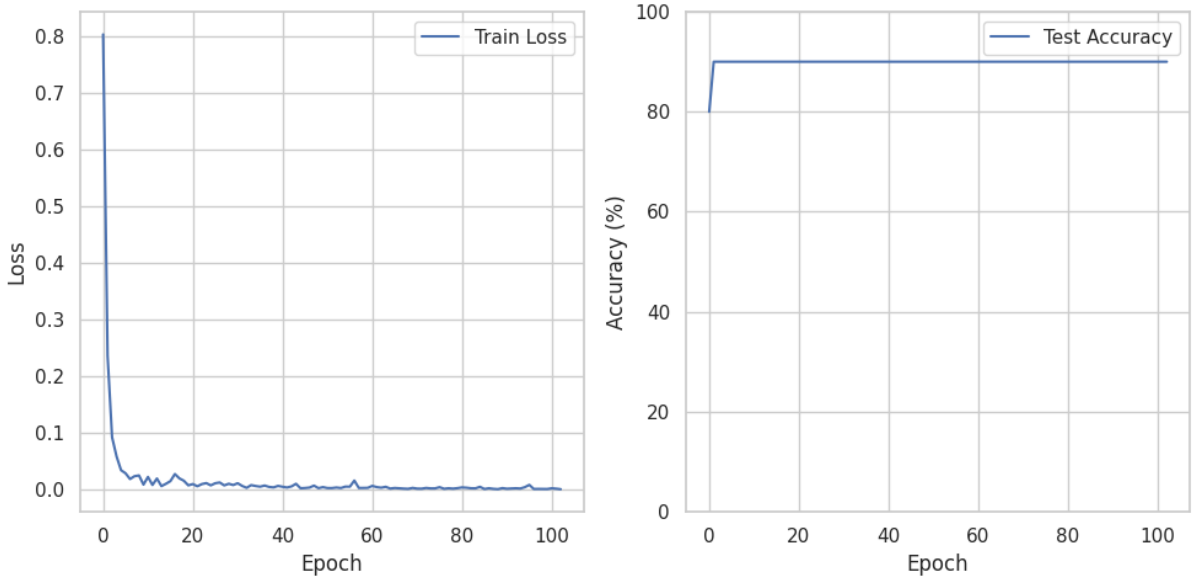

Figure 3 - Training loss and accuracy of the Vision Transformer

Here are the predictions made with the ViT model:

Field	Road