Arthur SATOUF 25/11/2023

Quick points/remarques

- The distribution of labels is unbalanced, leading to potential overfitting on the most representative class. Possible solutions include using WeightedRandomSampler to modify the loss function for each class. Other methods involve under-sampling roads or over-sampling fields, among others.
- Mislabeling has been observed during diagnostics. If possible, corrections should be made. However, in many cases, we have to accept them, especially in large real-world datasets where detection is challenging. In my case, no corrections were applied.
- Your labelled data has been placed in the "train_images" folder.
- A WebApp was developed for this challenge, and the DockerFile has been provided. You can check the GitHub repository for more details.

Data

Before proceeding with data balancing, processing, or transformation, it is essential to divide the data into three sets:

- Train: To train and prepare the model.
- Valid: To validate which model and strategies to choose.
- Test: For Stress test/final validation.(I didn't use your folder test_images)

Train 60% (“91 observations”), Valid 19% (“30 obs”) and Test 21% (”32 obs”)

Understanding the problem

Instead of immediately delving into sophisticated models, it's prudent to analyse what distinguishes an image as either a road or a field. As a human, I can discern the difference based on the shape of objects and the density of certain colours such as green and yellow (indicative of fields) and grey/black (indicative of roads). Analysing the histograms of a few images reveals that field and road images do not share the same density distribution. For field images, the peak hue density is more towards the end (high pixel values), while for road images, it is more towards the beginning (low pixel values).

Models

Baseline:

The baseline model employs a simple rule: If the mean of all pixel values in an image is below 128, it is classified as a field; otherwise, it is labelled as a road. The results indicate satisfactory performance for "roads" but reveal shortcomings in accurately classifying "fields," as illustrated in Table 1 below.

SVM Approach:

Instead of relying solely on the mean of the entire pixel, we consider both the mean and variance across the three colour channels. The results, as shown in Table 1, are highly promising, outperforming the baseline. Despite several rounds of hyperparameter optimization, further improvements seem elusive. It is crucial to closely monitor and consider these **key features: mean and variance**.

CNN Hybrid Approach:

Rather than building a CNN model from scratch, we leverage Transfer Learning. Specifically, we harness the GoogleNet architecture along with its pre-trained weights. This approach not only conserves