Problem

At Trimble/Bilberry, we continuously gather new images to upgrade our algorithms. As a Senior AI Engineer, given the constraint of only being able to annotate a limited number of images from a vast unlabelled dataset, how would you go about selecting the 1000 most promising images out of 1 million to enhance our algorithm through annotation ?

Solution

Quality Control: Begin by eliminating irrelevant or low-quality images from the dataset.

Rare or Critical Cases: Prioritize images that contain rare or critical situations that our model may struggle with, or cases that are crucial for the application domain.

Active Learning: Implement an active learning strategy by selecting the images on which the model is uncertain or has the highest prediction errors for the next round of annotation.

Cluster Analysis: Use unsupervised clustering techniques (e.g., k-means or hierarchical clustering) on the unannotated images to group similar images together. Select representative images from each cluster to ensure diversity.

Similarity Search: Extract pertinent features from the images. These features help gauge similarity between images in the dataset. Prioritize unannotated images that differ significantly from those currently being annotated.

Query by committee: Train a variety of models using the existing labeled data and have them collectively decide on predictions for unlabeled data. Label those points for which the "committee" exhibits the most disagreement.

In conclusion, the selection of the most promising images from a vast dataset should be a dynamic and iterative process that leverages active learning, diverse representation, and human expertise. By systematically combining these elements, we can make the most efficient use of our annotation resources to enhance our computer vision algorithms.