LIGHTLY

Dataset Filtering and Analytics Report

22/01/2021 10:50:35

LIGHTLY

General Information

Job Information

Metric	Value
Build Time	Fri Jan 22 09:49:11 UTC 2021
Sampling Method	Coreset Algorithm
Number of Images	7481
Number of Corrupt Images	0
Number of Duplicates	0
Number of Removed Images	3883
Number of Output Images	3598
Job Submitted	22/01/2021 10:46:45
Job Finished	22/01/2021 10:47:28
Total Processing Time	42s

Estimated Savings

Task	Annotation Savings*	CO2 Savings*
Image Classification	$ 2041.80	0.44 kg
Object Detection	$ 8167.20	1.63 kg
Semantic Segmentation	$ 40836.00	29.27 kg

*https://lightly.ai/report

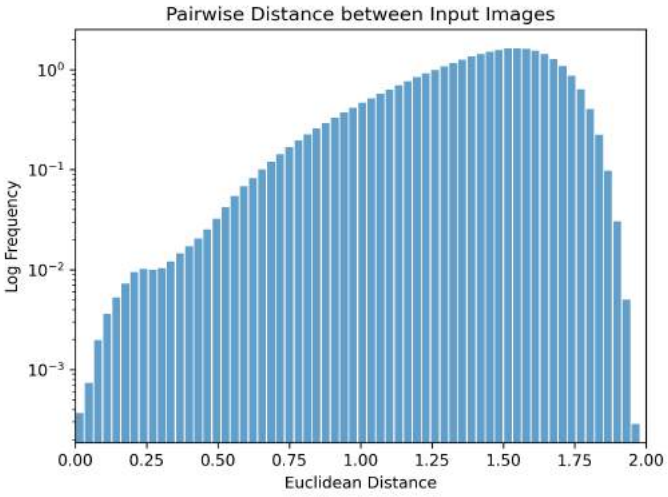
Statistics

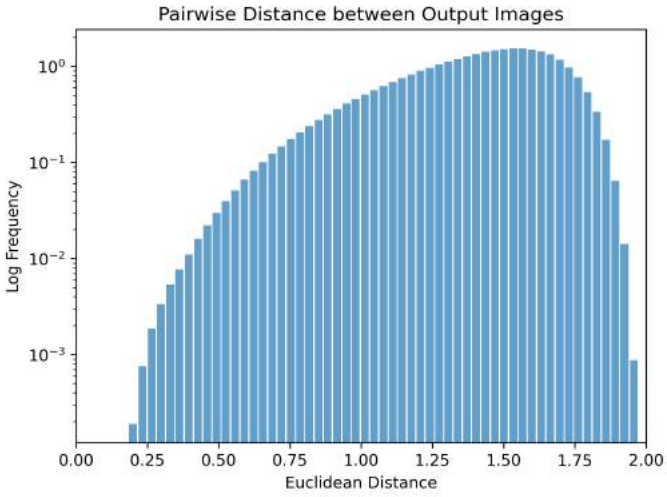
Metric	Before	After
Euclidean Distance (Mean)	1.3843	1.3849
Euclidean Distance (Min)	0.0000	0.1838
Euclidean Distance (Max)	1.9791	1.9734
Euclidean Distance (10th Percentile)	0.9952	0.9922
Euclidean Distance (90th Percentile)	1.6977	1.7069

Visualizations

Image Similarity in Input and Output Data

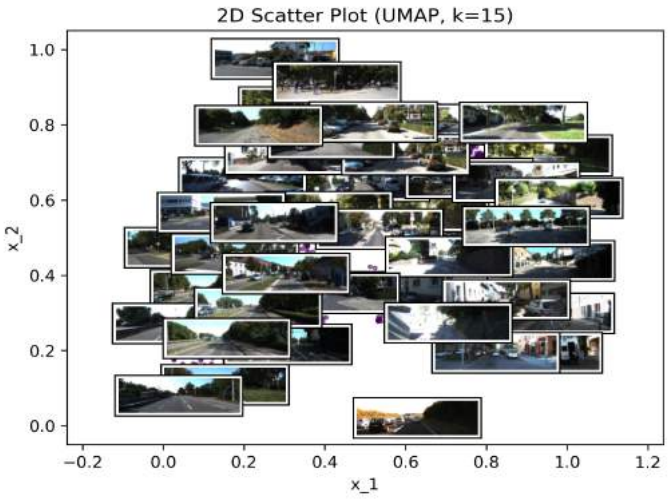
The plots below show the distribution of the pairwise distance between images in the input and output data. The histograms allow you to get information about the diversity of the dataset and whether the filter strength is well-chosen.

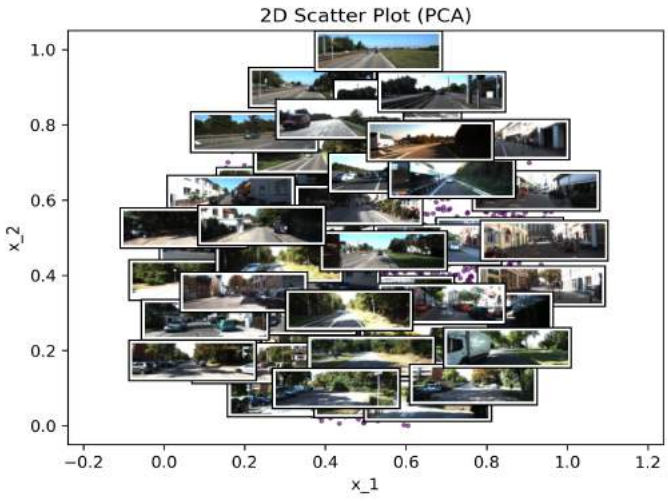




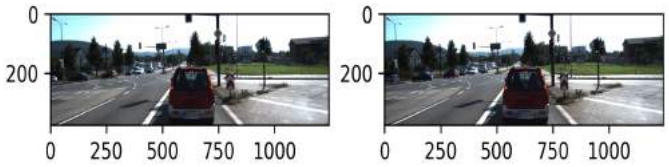
2D Scatter Plots of Output Data

Two-dimensional scatter plots help to understand the distribution of the data and may enable quick insights about outlier cases, dataset bias, or class imabalances.

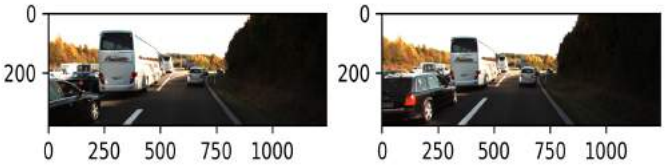




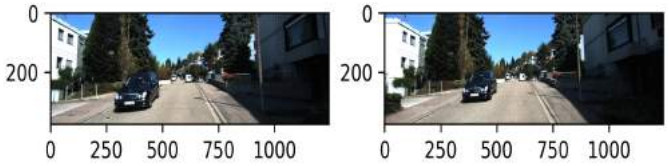
Retained (Left) and Removed (Right) Image with d = 0.01



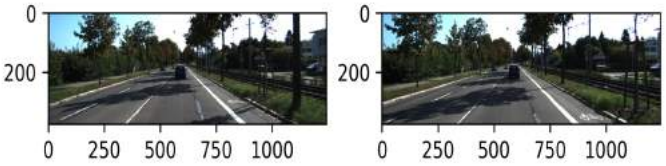
Retained (Left) and Removed (Right) Image with d = 0.13



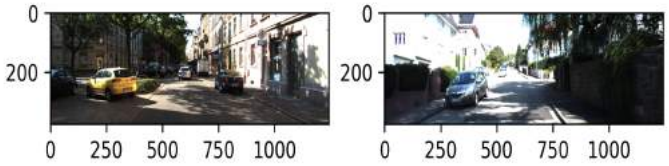
Retained (Left) and Removed (Right) Image with d = 0.17



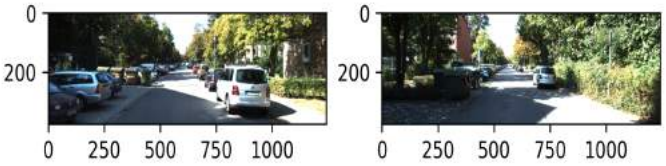
Retained (Left) and Removed (Right) Image with d = 0.20



Retained (Left) and Removed (Right) Image with d = 0.26



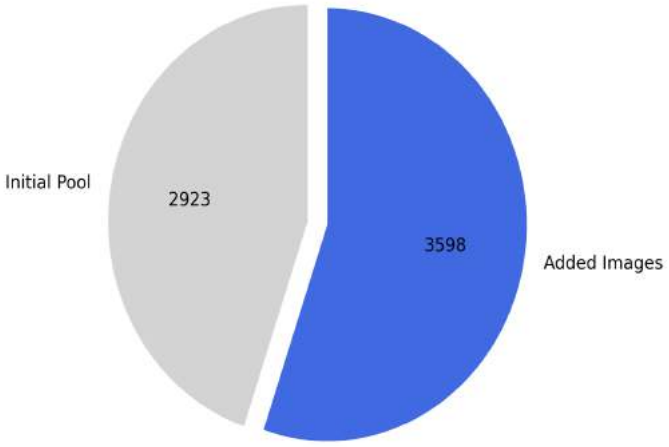
Retained (Left) and Removed (Right) Image with d = 0.32

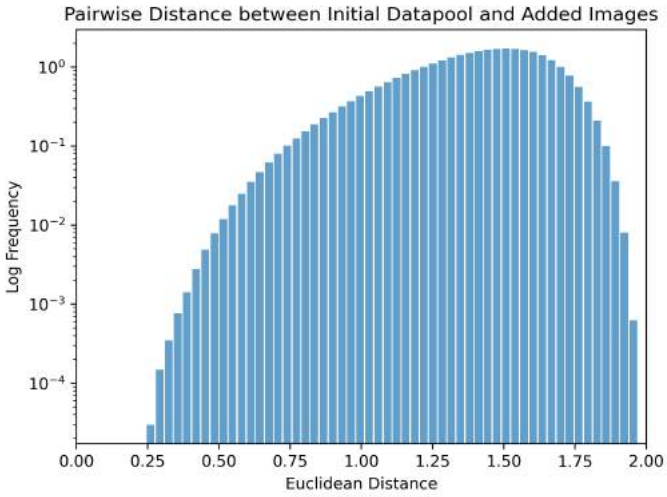


Datapool 1/2

Proportion of Added Images and Pairwise Distances

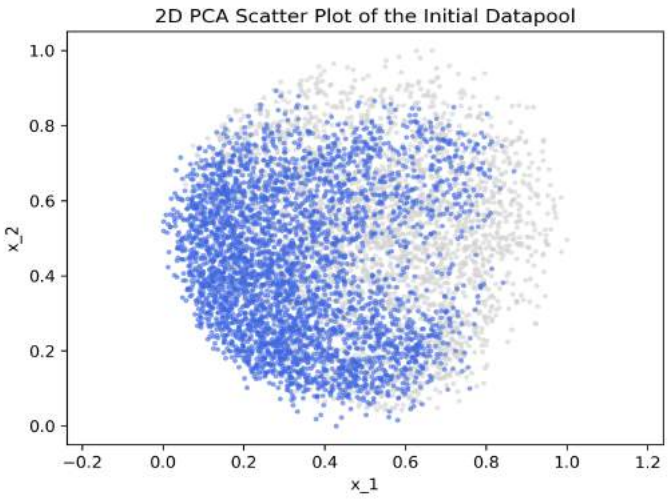
The figures below help to understand how many new images were added to the datapool and how similar the new images are to the ones selected in previous iterations.

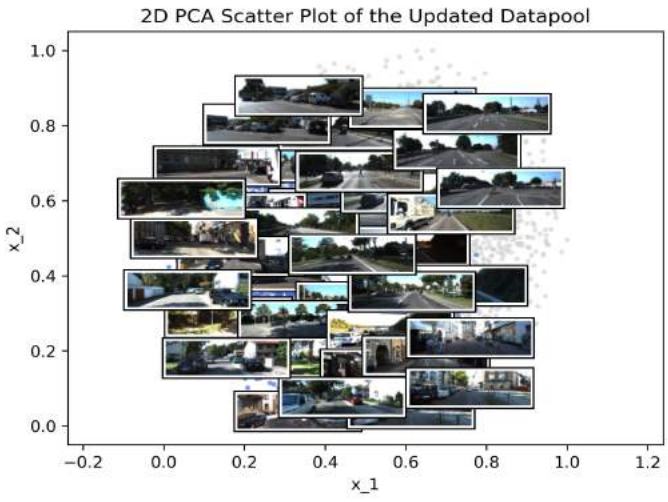




2D Scatter Plots of the Datapool (PCA)

The two-dimensional scatter plots of the datapool give an overview over the images which were added to it.





Datapool 2/2

2D Scatter Plots of the Datapool (UMAP, k=15)

The two-dimensional scatter plots of the datapool give an overview over the images which were added to it.