



Dataset Purification and Elevation Report


14-Mar-2020--10:28:40

General Information

Job Information

Metric	Value
Glob expression	*.jpg
Number of Samples	118287
Job Submitted	14-Mar-2020--08:58:38
Job Finished	14-Mar-2020--10:28:40
Total Processing Time	1h 30m 2s

Estimated Savings

Metric	Annotation Savings*	CO2 Savings* 
Image Classification	\$ 4 533	5.89 kg
Object Detection	\$ 18 132	21.76 kg
Semantic Segmentation	\$ 90 661	389.84 kg

*<https://www.whattolabel.com/report>

Statistics

Metric	Value
Number of Removed Samples	90661
Min. Distance	2.9882
Max. Distance	51.4302

The algorithm examines the distance in the vector space.
Redundant data points lie close to each other.

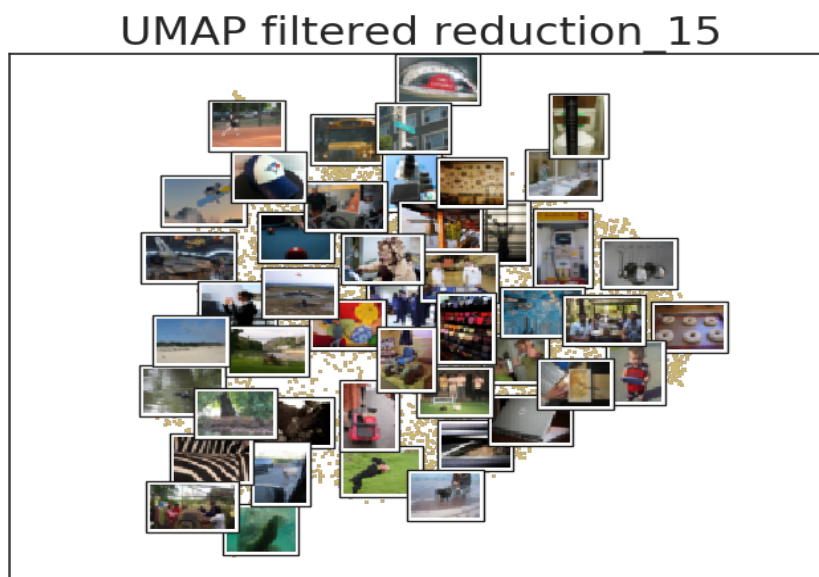
Min. Percentile	15.9212
Max. Percentile	26.8898

Indicates the distance at which 5% of the data (min. percentile)
is below or 95% of the data (max. percentile) is below this distance.

Statistics

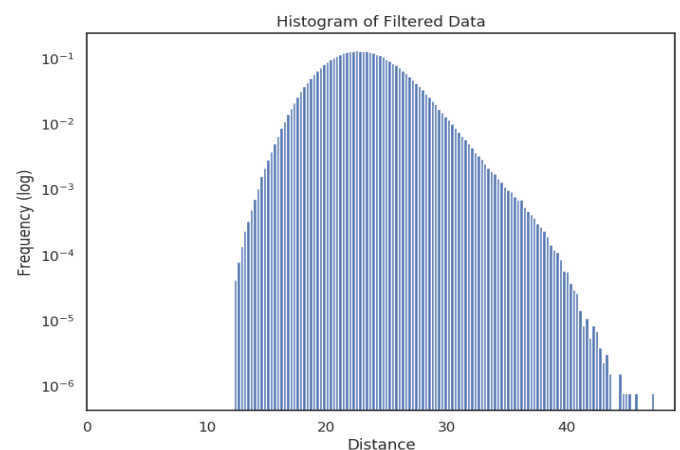
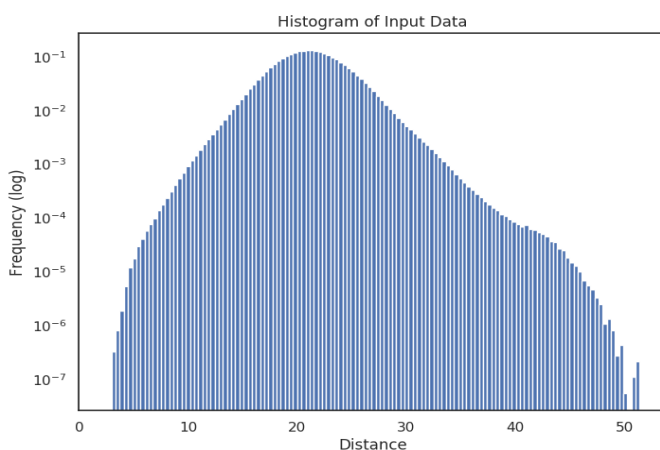
UMAP Reduction 15

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualization similarly to t-SNE, but also for general non-linear dimension reduction. We run the UMAP algorithm on the dataset. This might help you formulate a quick hypothesis about the dataset.



Histogram of Input and Output Data

We show you the histograms of distances within our neighborhood graphs we build for filtering the data. The histograms allow you to get information such as the diversity of the dataset and whether the filter strength is chosen correctly.



Examples of Removed Data

The left side shows the removed picture while the right side shows the retained picture.

