T-MARS: Improving Visual Representations by Circumventing Text Feature Learning

Carnegie Mellon University

Pratyush Maini †* Sachin Goyal †*

Zachary C. Lipton † Zico Kolter †✦ Aditi Raghunathan †

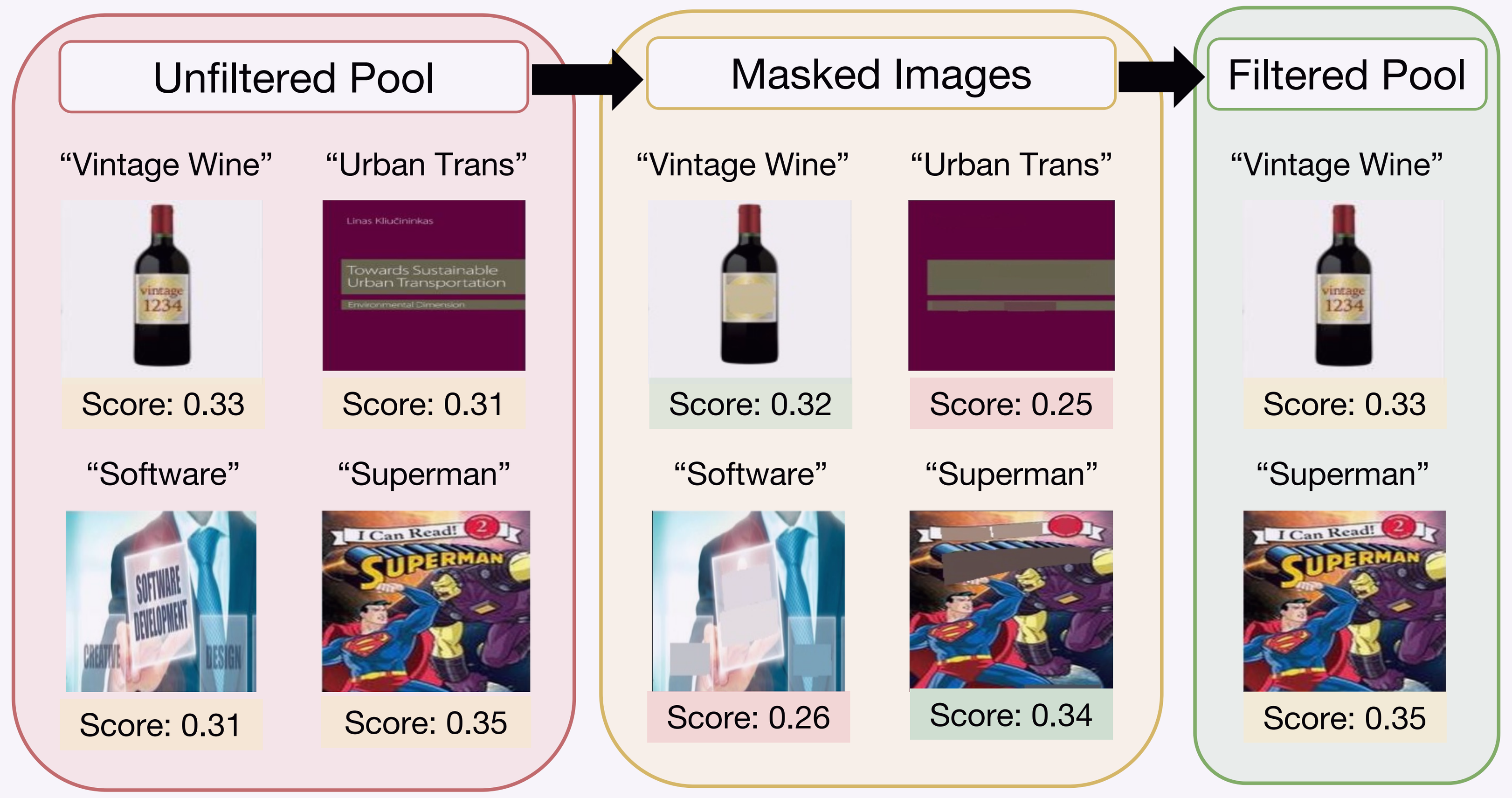
*Equal Contribution | † Carnegie Mellon University | ✦ Bosch Center for AI

TLDR: We filter web-scale datasets used for training CLIP to learn better visual representations and achieve state-of-art zero-shot accuracy on vision tasks.

A closer look at Web Data

- Web-images contain *text* inside them.
- Often, the *text* is the only feature correlated with the caption (Category 5).
- Such images promote the model to learn OCR and not visual representations

T-MARS for Web Data Curation



T-MARS is based on filtering out images dominated by text features.

1. **Text Detection**: Perform text detection using an off-the-shelf OCR model.
2. **Text Masking**: In-paint the pixels where text is detected with average nearby pixel value.
3. **Re-scoring & Filtering**: Retain images whose corresponding *masked* images have a high CLIP similarity score with the original caption, i.e. have visual features correlated with the caption.

Other Contributed Baselines

We also propose 2 approaches drawing insights from the literature on hard example mining:

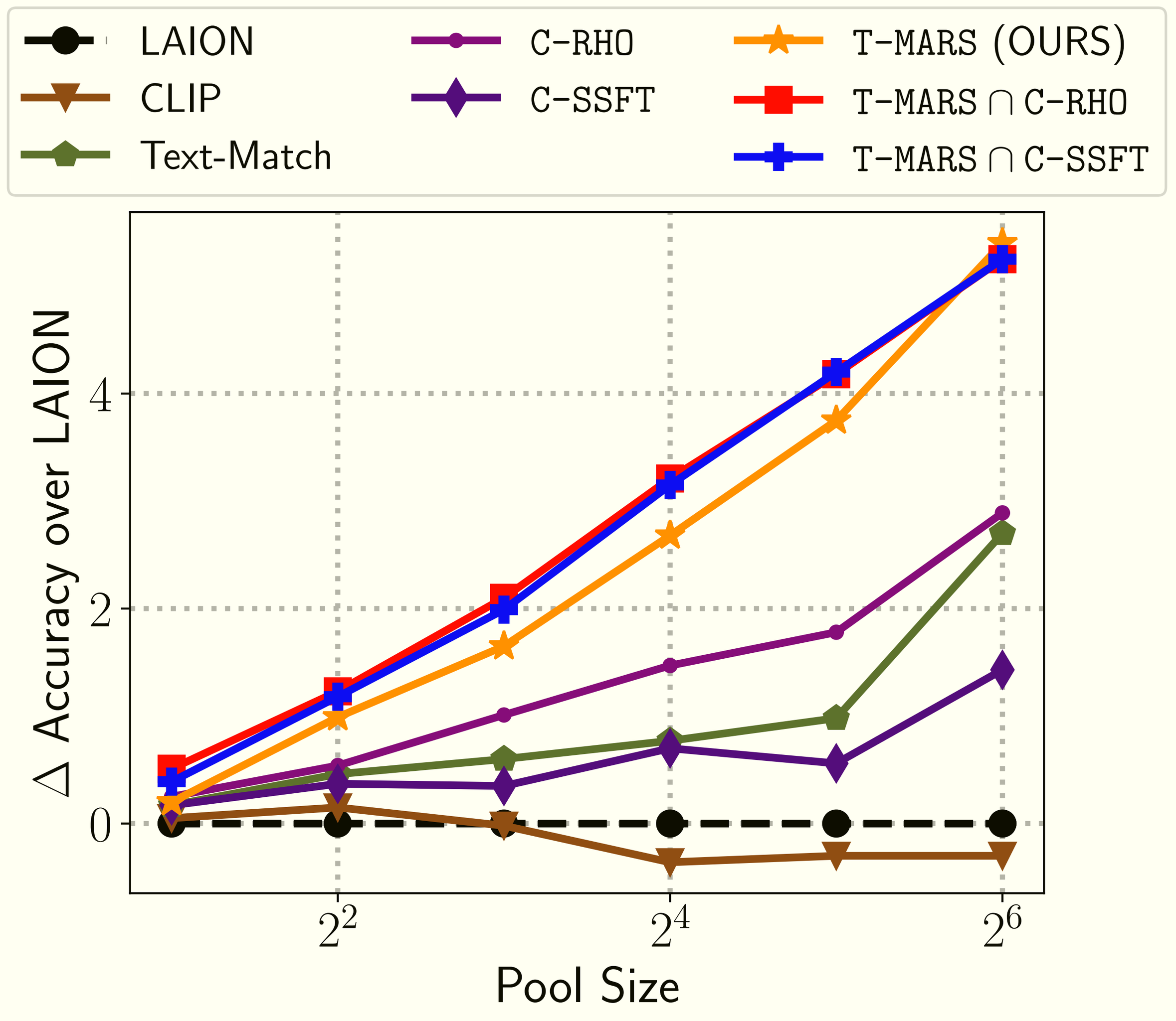
- C-SSFT: Identify mislabeled examples based on change in CLIP score when finetuning a base scoring model on a held-out set
- C-RHO: Prioritize training on samples with low validation model loss but high training loss.

SOTA on DataComp

Filtering	Dataset size	ImageNet	ImageNet dist. shifts	VTAB	Retrieval
	medium (128M)				
No filtering	128M	17.6	15.2	25.9	17.4
Basic Filtering	30M	22.6	19.3	28.4	19.2
LAION filtering	13M	23.0	19.8	30.7	17.0
CLIP score (L/14 30%)	38M	27.3	23.0	33.8	18.3
T-MARS	25M	33.0	27.0	36.3	22.5
T-MARS ∩ C-RHO	15M	30.3	24.9	34.9	19.9
T-MARS ∩ C-SSFT	23M	**33.8**	**27.4**	**37.1**	**23.1**

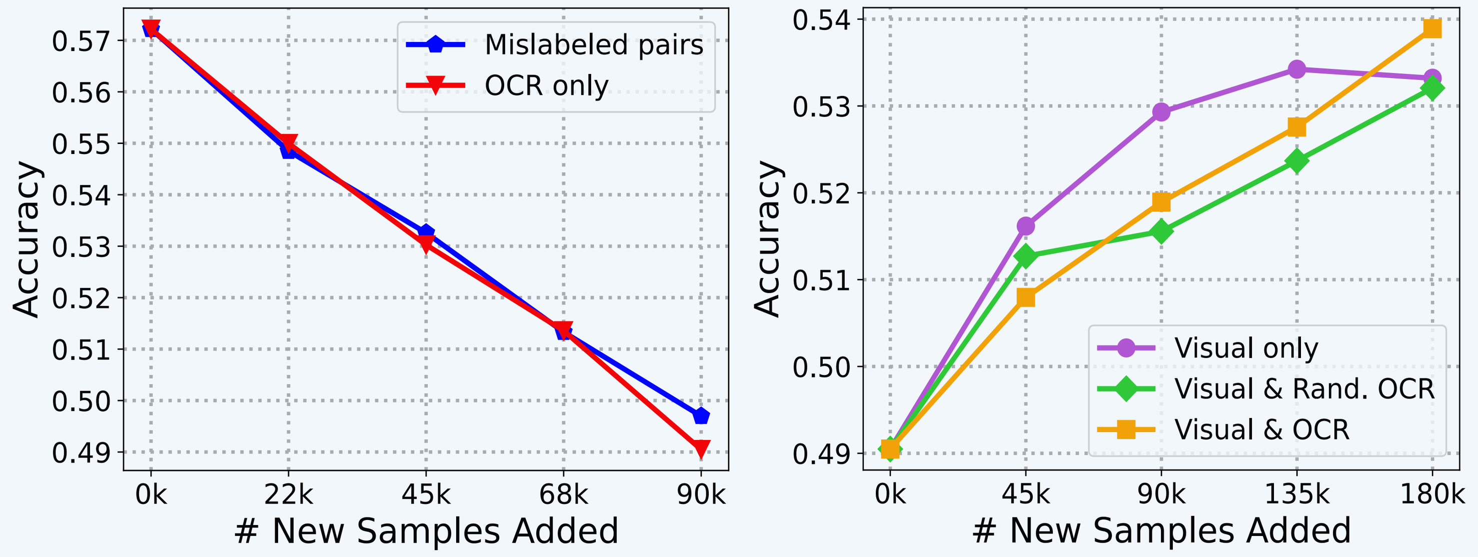
- T-MARS outperforms the top of the leaderboard on DataComp (a data filtering benchmark) by 6.5% on ImageNet.

Scaling Trends



- Accuracy gains linearly increase as data and compute double from 2M to 64M samples from the LAION dataset.

Utility of Various Data Categories



- Images with *text* as the *only* predictive feature hurt as much as adding *mislabeled* examples to the dataset.
- Images with both *visual* & *text* features are as useful as those with *no text* & should not be removed from the dataset.
- With the ML community focused on scaling up datasets, this shows that pruning off *'bad data'* can have 3× more utility than adding more *'good'* samples.

New work on Diminishing Utility of Different Data

Scaling Laws for Data Filtering -- Data Curation cannot be Compute Agnostic (Best Paper Award at DPFM ICLR Workshop)

TLDR: High-quality data is limited and loses utility with repetitions. So how to determine the optimal data curation strategy → scaling laws for web data curation!!