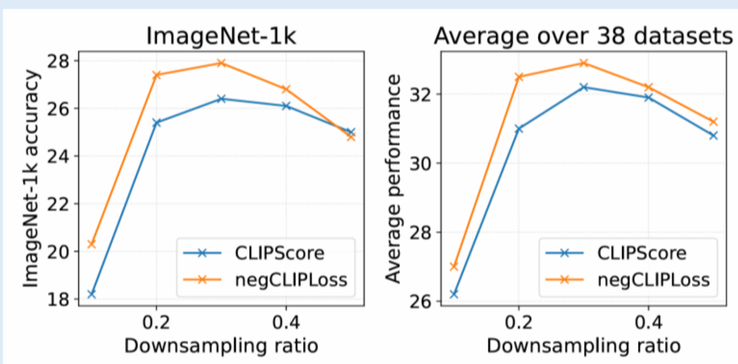
CLIPLoss and Norm-Based Data Selection Methods for Multimodal Contrastive Learning

Yiping Wang[1,*], Yifang Chen[1,*], Wendan Yan[1,3], Alex Fang[1], Wenjing Zhou[2], Kevin Jamieson[1], Simon Shaolei Du[1]

[1] University of Washington, [2]University of Michigan, [3]Microsoft Corporation, [*]Equal Contribution

Background

- Data selection is critical for large-scale multimodal pretraining, particularly for noisy web-curated datasets.
- Current best data selection approaches for CLIP pretraining leverage external non-CLIP models [2,3] (like BLIP and OCR models, etc) or the external large-scale high-quality pretraining datasets (like HQITP-350M[1]), while the potential of CLIP embedding is under-explored.



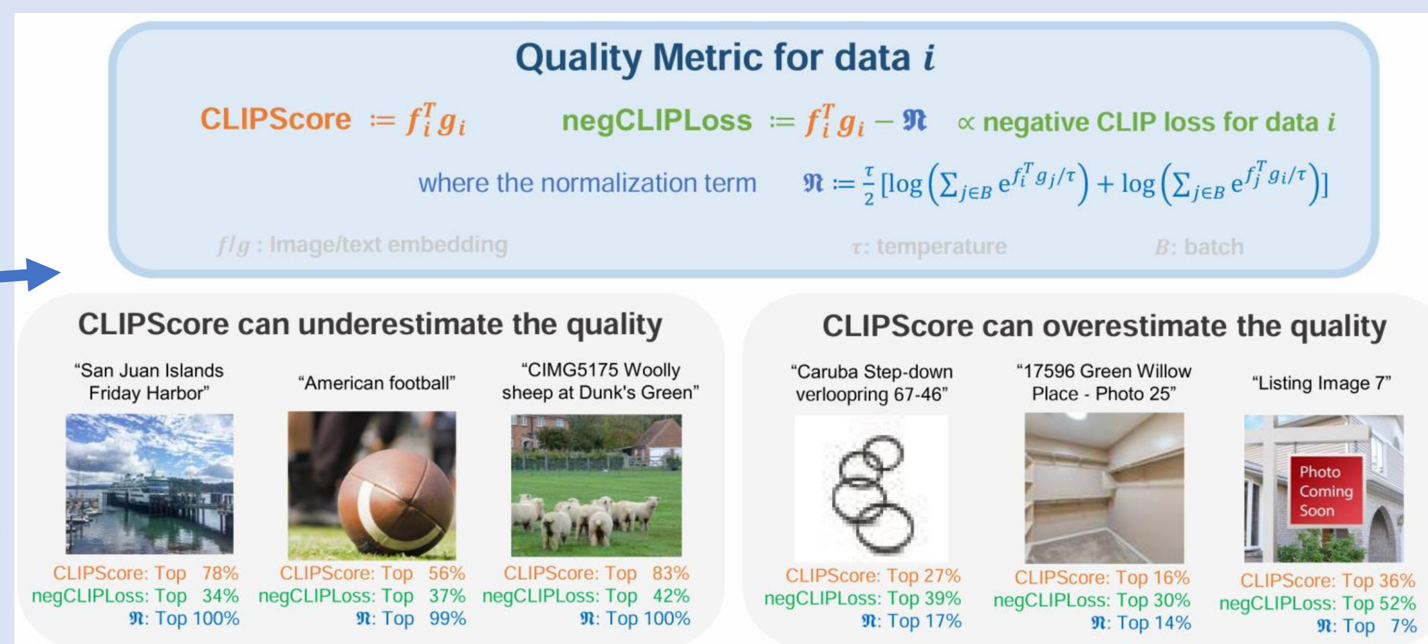
Key Takeaway

- **CLIP Loss is better than CLIPScore in image-text data selection.** Simply replacing CLIPScore with negative CLIP Loss (negCLIPLoss) can consistently produce better quality measurement in data selection for CLIP pretraining.
- **New SOTA on DataComp-medium.** Our methods can achieve a new SOTA on the DataComp benchmark by combining them with the current top techniques.
- **High Efficiency and Universality.** Our methods don't need external non-CLIP models or the external large-scale pretraining datasets as in previous works. They are very efficient and also universal to different CLIP teacher models.

Code: https://github.com/ypwang61/negCLIPLoss_NormSim
DataComp samples and checkpoint:
https://huggingface.co/ypwang61/negCLIPLoss_NormSim/tree/main/medium_scale
DataComp Benchmark: https://www.datacomp.ai/dcclip/leaderboard.html

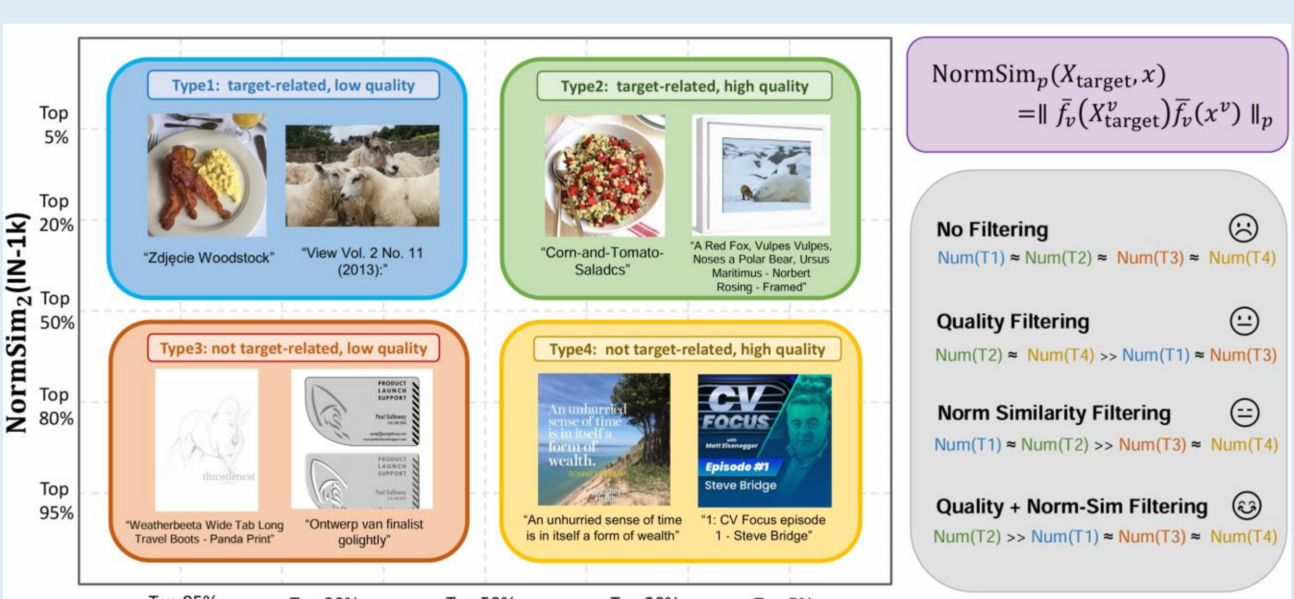
negCLIPLoss ourperforms CLIPScore

- The difference between CLIPScore (similarity of one image-text pair) and the negative of CLIP Loss (negCLIPLoss) is only a normalization term.
- This term calculates the similarity among **cross**-image-text pairs, which **penalizes the image/text features shared by different data** (Like monotonous patterns/colors in images, or "Photo"/"Image" in texts), and rewards more distinctive data.
- Compared with CLIPScore, negCLIPLoss consistently provides better estimates of data quality.



NormSim: Estimate source-target similarity

- Quality and source-target similarity are orthogonal metrics. OCR-style image-text pairs can also have high-quality scores.
- Using p-norm similarity on the image embeddings to evaluate the source-target relevance. We use the validation data from the 24 downstream tasks as target. When $p = 2$, it's theoretically provable that $\mathrm{NormSim}_2$ is optimal under a linear setting. And in experiments, $p = \infty$ has the best performance.
- We also have a variant $\mathrm{NormSim}_2$-D (Dynamic) which is target-data-free.



(a) Data Type for Different **negCLIPLoss** and $\mathbf{NormSim_2}$

(b) Data Selection Methods Comparison

Experiment Results

Setup: Following **DataComp-medium** pipeline. It contains 128M low-quality data for filtering (successfully download 110M). After obtaining subsets with some data filtering strategies, it will train a CLIP-B/32 model with a fixed budget.

E1: methods Just utilizing CLIP embedding

Strategy	IN-1k	VTAB	Avg.
CLIPScore	26.4	32.6	32.2
negCLIPLoss **(Ours)**	27.9	33.2	32.9
negCLIPLoss ∩ $\mathrm{NormSim}_2$-D **(Ours)**	<u>29.8</u>	<u>34.8</u>	<u>34.1</u>
negCLIPLoss ∩ $\mathrm{NormSim}_\infty$ **(Ours)**	**31.7**	**36.0**	**35.0**

E2: Comparing All methods

Strategy	IN-1k	VTAB	Avg.
DFN [1]	36.0	36.2	35.4
Devil [2]	31.0	35.9	34.5
DFN ∪ HYPE [3]	<u>36.4</u>	<u>38.5</u>	36.8
DFN ∪ **Ours**	<u>36.4</u>	**38.6**	<u>37.6</u>
DFN ∪ HYPE ∪ **Ours**	**37.3**	<u>38.5</u>	**37.7**

E3: Universality : Our methods applied for OpenAI's CLIP-B/32, CLIP-L/14, and the public version of DFN models.

More details in our papers. In Leadboard we are applying our methods on the complete data pool (i.e., 128M data pool)

[1] Fang, A., Jose, A. M., Jain, A., Schmidt, L., Toshev, A., & Shankar, V. (2023). Data filtering networks. *arXiv preprint arXiv:2309.17425*.
[2] Yu, H., Tian, Y., Kumar, S., Yang, L., & Wang, H. (2023). The devil is in the details: A deep dive into the rabbit hole of data filtering. *arXiv preprint arXiv:2309.15954*.
[3] Kim, W., Chun, S., Kim, T., Han, D., & Yun, S. (2024). HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts. arXiv preprint arXiv:2404.17507.