

Effective pruning of web-scale datasets based on complexity of concept clusters



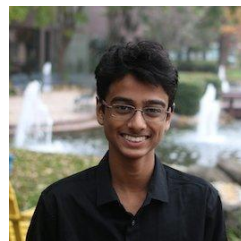
Amro Abbas*



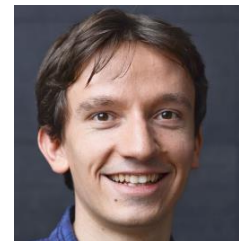
Evgenia Rusak*



Kushal Tirumala



Wieland Brendel



Kamalika Chaudhuri

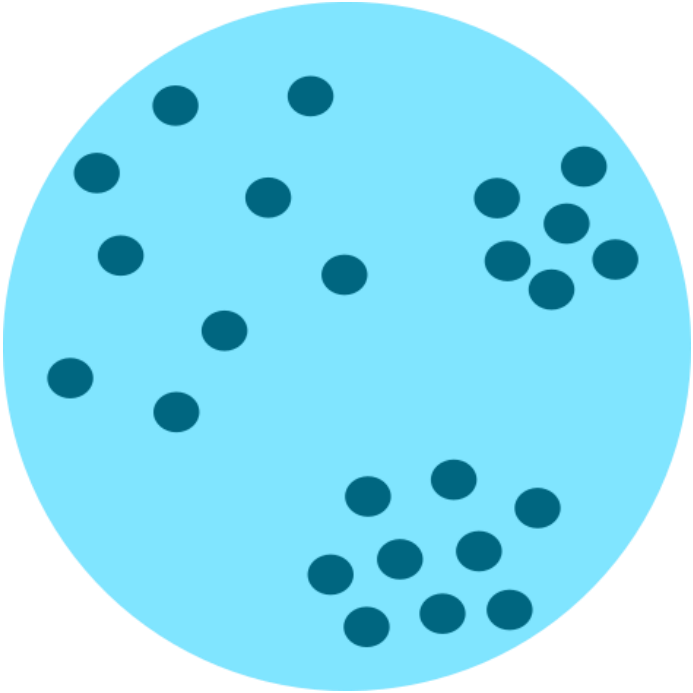


Ari Morcos



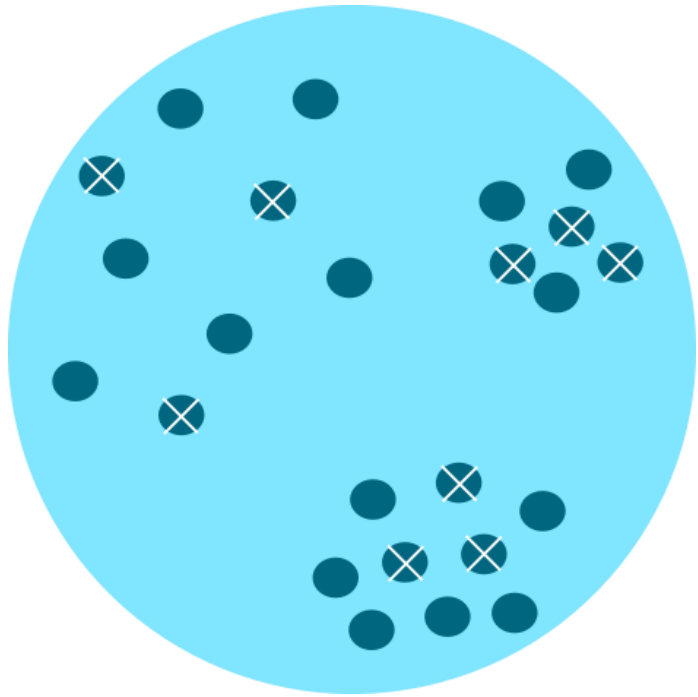
Task: Improve training efficiency

Dataset

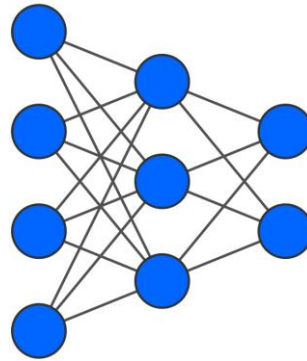


Task: Improve training efficiency

Dataset



train a model

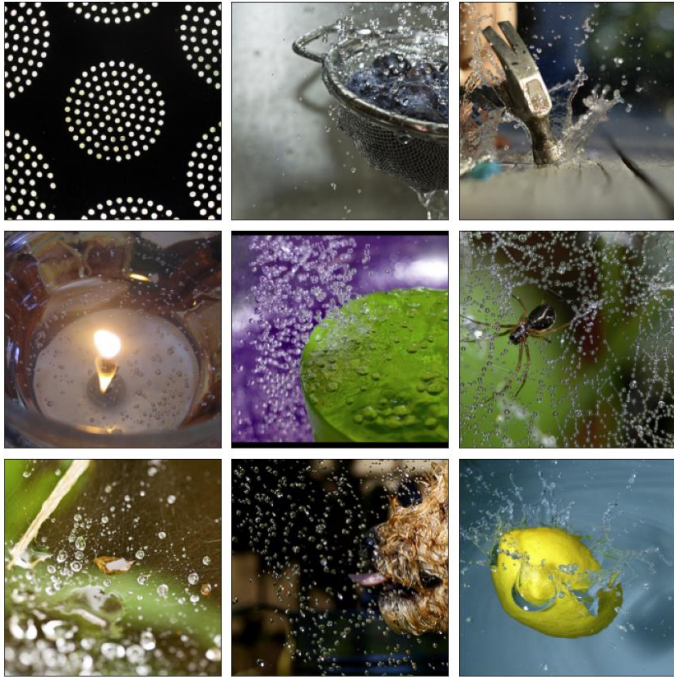


classify

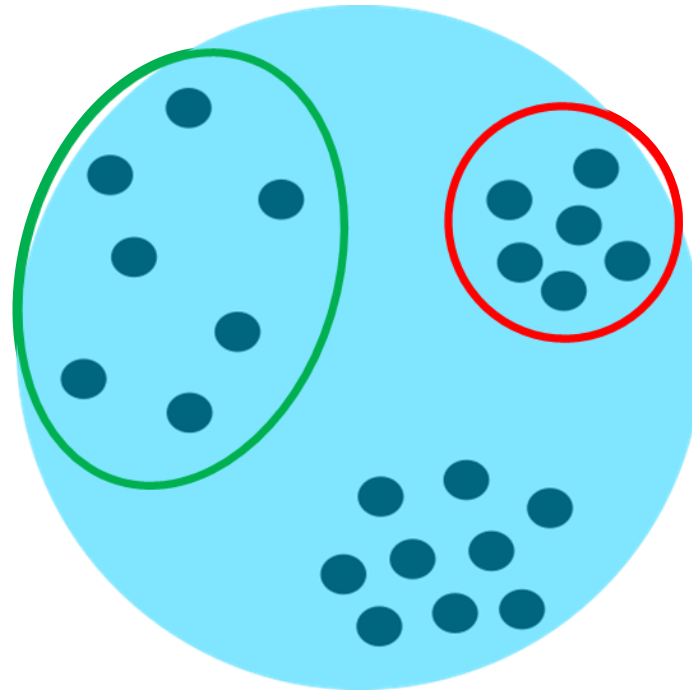


How many examples do we need to learn a concept?

Cluster A



Dataset

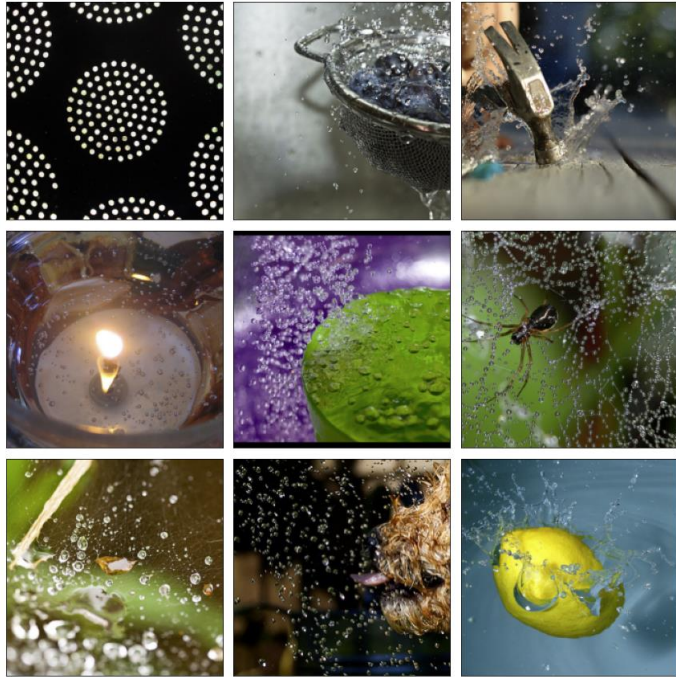


Cluster B

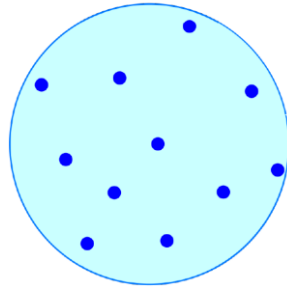


How many examples do we need from each cluster?

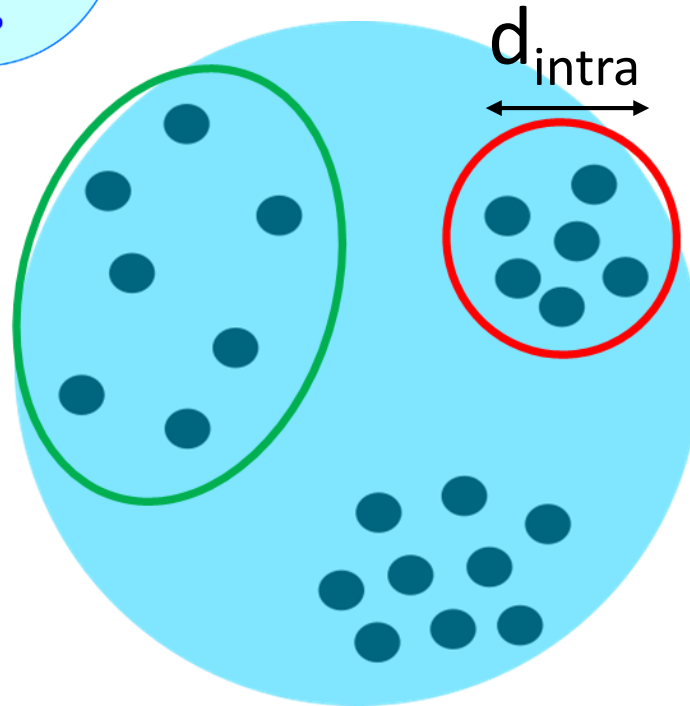
Cluster A



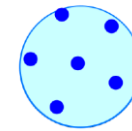
high d_{intra}



Dataset



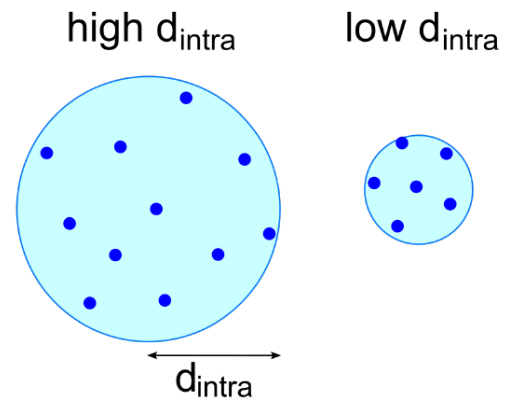
low d_{intra}



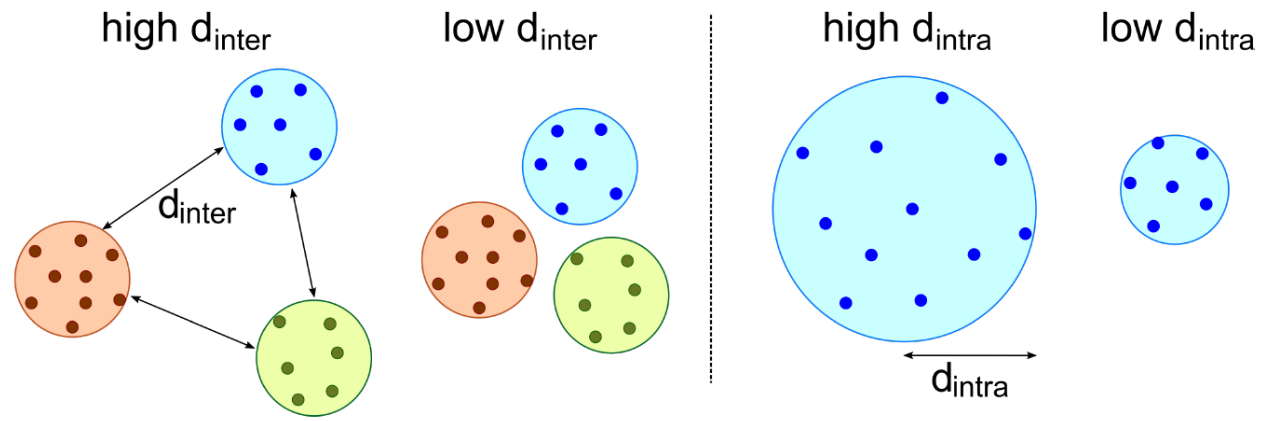
Cluster B



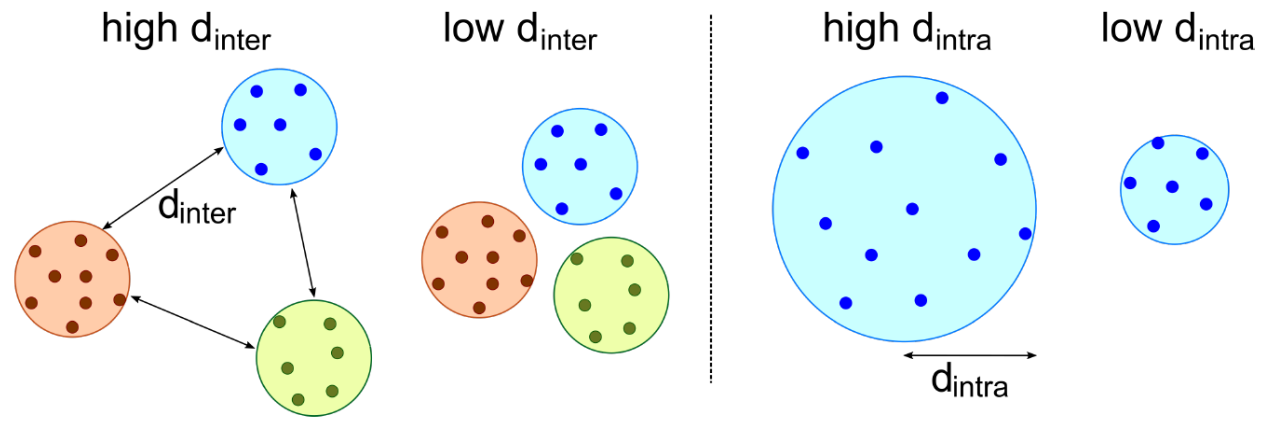
Density-based pruning



Density-based pruning



Density-based pruning



$$C_j = d_{inter,j} \cdot d_{intra,j}$$

$$P_j = \frac{\exp(C_j/\tau)}{\sum_i^n \exp(C_i/\tau)}$$

Density-based pruning: Method summary

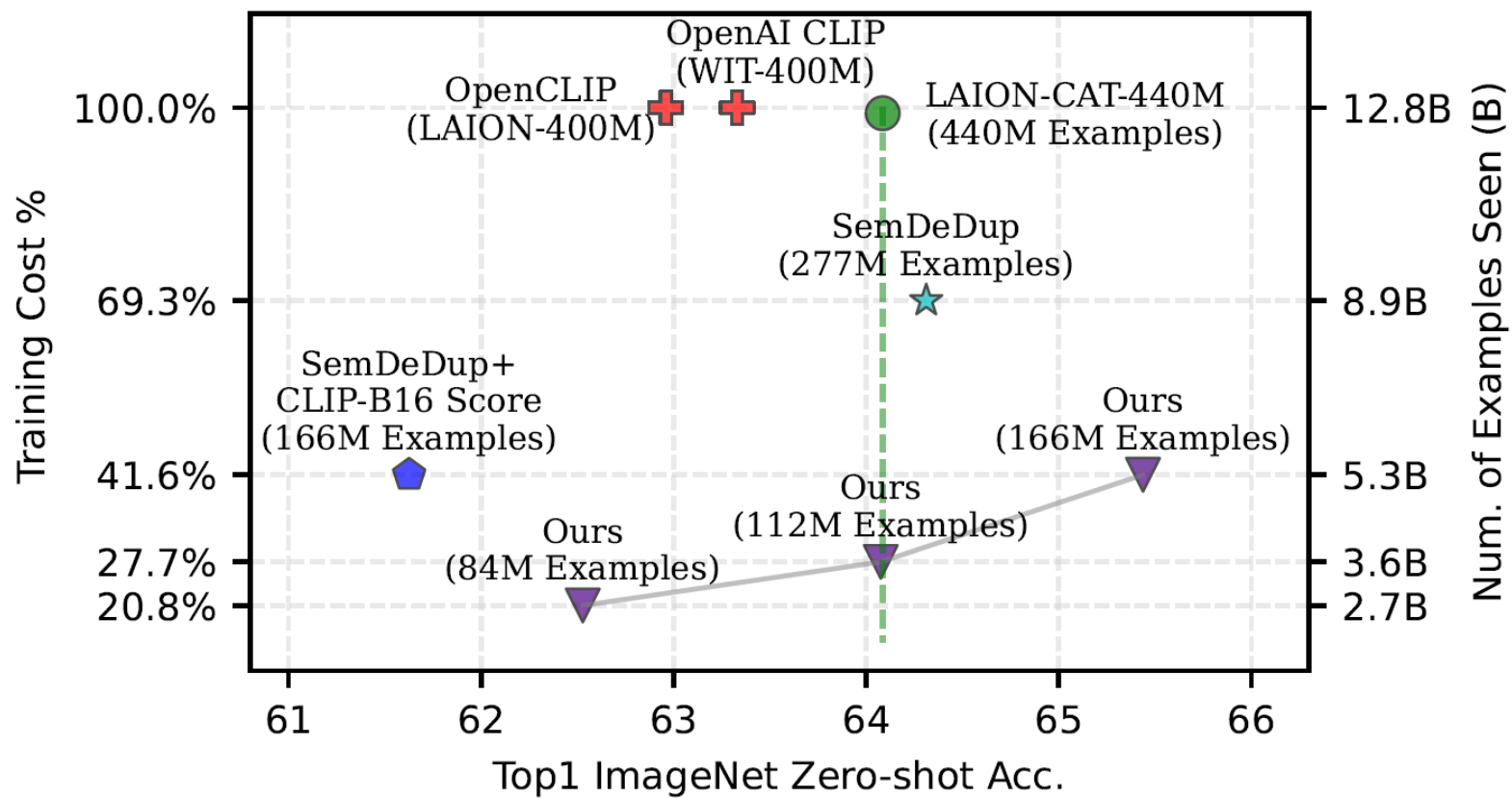
1. Deduplicate LAION / DataComp.
2. Calculate image embeddings with a pretrainer encoder.
3. Cluster the embeddings with kmeans.
4. Calculate d_{intra} and d_{inter} for each cluster.
5. Calculate the number of examples per cluster N_j .
6. Keep the furthest N_j samples from each cluster centroid.

Results on DataComp Medium

Task: Train a CLIP model on 128M examples seen.

Method	Size	ImageNet	ImageNet dist. shifts	VTAB	Retrieval	Average
TMARS	25M	33.00	27.00	36.30	22.50	36.10
Image-based \cap CLIP Score (L/14 top 30%)	14M	29.70	23.90	34.60	23.10	32.80
CLIP Score (L/14 top 30%)	38M	27.30	23.00	33.80	25.10	32.80
Ours (DeDup, 80% + CLIP-L/14 Score, 50% + DBP)	19.2M	33.35	24.73	37.26	26.82	34.52
Ours (DeDup, 80% + CLIP-L/14 Score, 40% + DBP)	19M	32.02	25.74	37.26	26.80	35.35

Results on LAION-400M



Key message: Taking the complexity of different concepts into account when designing pruning methods can reduce redundancy and improve training efficiency.

Path forward: Combine different pruning techniques.

Thank you for your attention 😊.

Meet me at my poster!



Concept-based data-driven curation of large-scale datasets



Amro Abbas¹ Evgenia Rusk^{2,3,4} Wieland Brendel^{5,6,7} Kamalika Chaudhuri⁸ Ari Morcos⁹

¹Work done during a residency at Meta AI ²Work done during an internship at Meta AI ³University of Tübingen, Germany ⁴International Max Planck Research School for Intelligent Systems ⁵Max-Planck Institute of Intelligent Systems, Germany ⁶ELLIS Institute Tübingen ⁷Tübingen AI Center ⁸Meta AI ⁹Datology AI ^{*}equal contribution

✉ amroabbas@gmail.com ✉ evgenia.rusk@uni-tuebingen.de

Task: Improve Training Efficiency

Key message: Taking the complexity of different concepts into account when designing pruning methods can reduce redundancy and improve training efficiency.

Self-Supervised Prototypes Pruning

Source: et al., "Beyond neural scaling laws: beating power-law scaling via data pruning"

Scaling Self-Supervised Prototypes Pruning to LAION

Abbas et al., "SelfDeDup: Data-efficient learning at web-scale through semantic deduplication"

How many examples do we need from each cluster?

Density-based pruning (DBP)

$$C_j = d_{\text{inter},j} \cdot d_{\text{intra},j}$$

$$P_j = \frac{\exp(C_j/r)}{\sum_i \exp(C_i/r)}$$

Density-based pruning (DBP)

Method summary:

1. Deduplicate LAION / DataComp.
2. Calculate image embeddings with a pretrainer encoder.
3. Cluster the embeddings with kmeans.
4. Calculate d_{inter} and d_{intra} for each cluster.
5. Calculate the number of examples per cluster N_j .
6. Keep the furthest N_j samples from each cluster centroid.

Results on DataComp Medium

Method	Size	ImageNet	ImageNet dist. shifts	VTAB	Retrieval Average	
TMARS	25M	33.00	27.00	36.30	22.50	36.10
Image-based \cap CLIP Score (L/14 top 30%)	14M	29.70	23.90	34.60	23.10	32.80
CLIP Score (L/14 top 30%)	38M	27.30	23.00	33.80	25.10	32.80
Ours (DeDup, 80% + CLIP-L/14 Score, 50% + DBP)	19.2M	33.35	24.73	37.26	26.82	34.52
Ours (DeDup, 80% + CLIP-L/14 Score, 40% + DBP)	19M	32.02	25.74	37.26	26.80	35.35

Deduplication and CLIP score filtering are important.

Deduplicate?	DBP	Dataset Size (M)	ImageNet top-1 acc.	Average acc.
No	No	26.76	27.48	
No	Yes	20.34	29.91	
Yes	No	19.40	26.10	
Yes	Yes	19.06	30.96	

