DATACOMP: In search of the next generation of multimodal datasets.

What’s next?

Чуканов Тимофей Вячеславович, БПМИ201

27.11.2023

Filtering track. Medium

Rank	Created	Submission	ImageNet acc.	Average perf.
1	11-08-2023	Hype sampler + DFN	0.382	0.379
2	11-07-2023	Hype sampler	0.346	0.373
3	10-02-2023	Data Filtering Networks	0.371	0.373
4	09-08-2023	The Devil Is in the Details	0.320	0.371
5	09-08-2023	TMARS + SSFT	0.338	0.362
6	08-17-2023	T-MARS: Improving Visual Representations by Circumventing Text Feature Learning	0.330	0.361
7	09-08-2023	The Devil Is in the Details - ImageNet best	0.336	0.355
8	08-25-2023	SIEVE	0.303	0.354
9	09-05-2023	Density-based Self-supervised Prototypes Pruning	0.334	0.345
10	09-07-2023	OCR and Naive english filtering	0.294	0.343
11	08-22-2023	WS (baselines)	0.305	0.342
12	07-26-2023	Mixed rules	0.303	0.337
13	04-28-2023	Baseline: Image-based ∩ CLIP score (L/14 30%)	0.297	0.328

The Devil is in the Details: A Deep Dive into the Rabbit Hole of Data Filtering (Yu H. et al. 2023)

Filtering track. Top-4, Top-7 M.

Single-modality filtering:

- Дедупликация посредством KNN.
- Удаление частых и некачественных текстов. Оставляют только тексты на английском.
- Удаление фото, где лица очень большую площадь занимают, либо соотношение сторон вне (0.33, 3.33).

Cross-modality filtering:

- Flipped image CLIP score + BLIP.

Distribution alignment:

- Дублируют ‘хорошие’ пары image-text.
- Удаляем похожие пары внутри кластеров, чтобы было примерно одинаково число объектов на кластер.
- Находят изображения с числами и добавляют их в итоговый датасет.

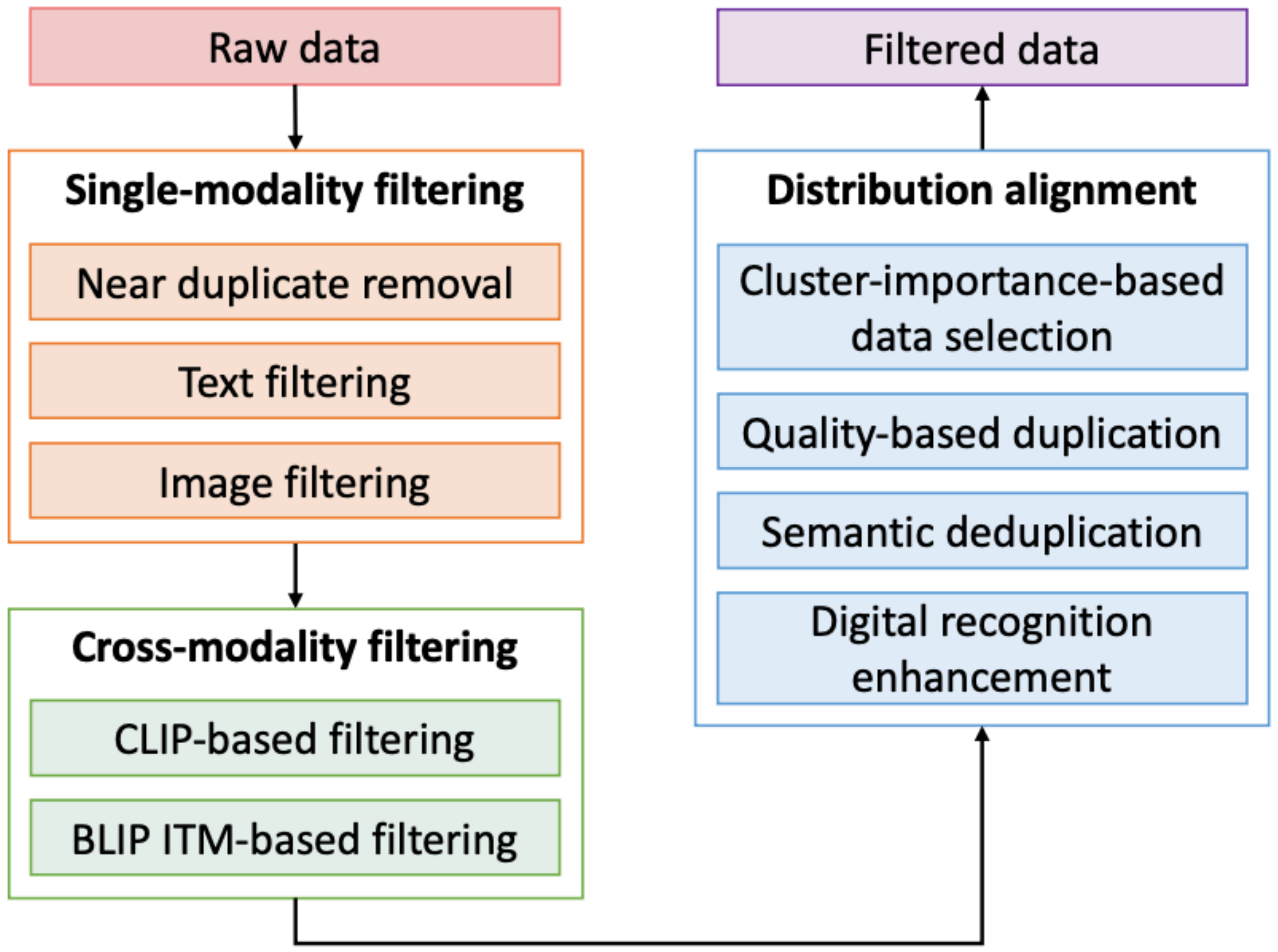


Рис. 1

DFNs: Data Filtering Networks (Fang A. et al. 2023)

Filtering track. Top-1 L, XL. Top-3 M.

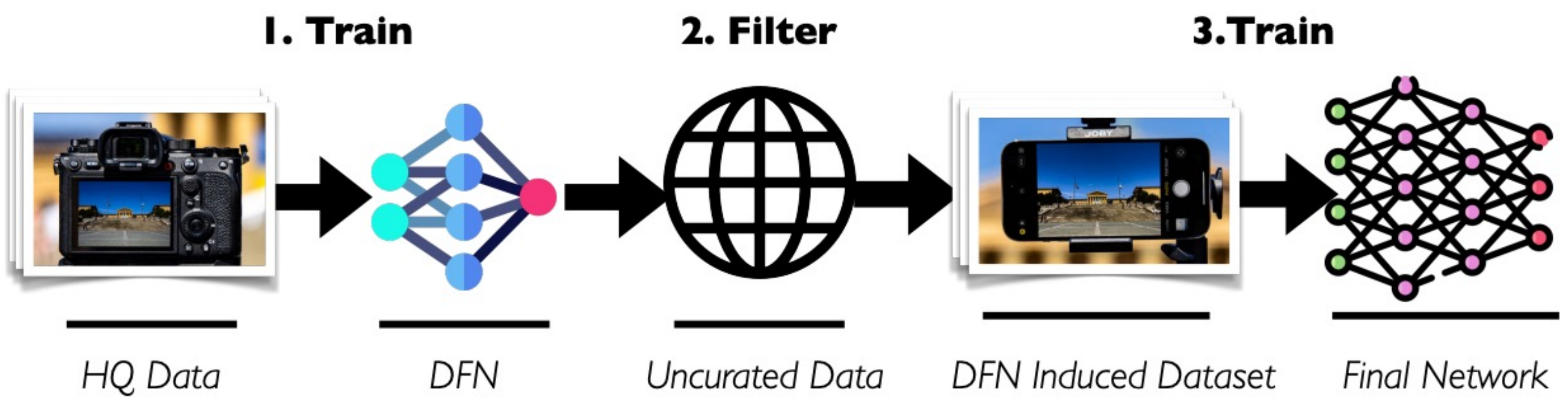


Рис. 3

Качество данных для обучения DFN влияет на качество конечной модели. (Рис. 2)

Архитектура CLIP наилучшим образом подходит для DFN. (Таблица 1):

- **Binary Filter: Filter Dataset = positives, Common Crawl = negatives.**
- **M3AE: reconstruction loss в качестве критерия фильтрации.**

DFN Type	Filter Dataset	ImageNet	Average
No Filter Baseline	None	0.176	0.258
ResNet-34 Image Binary Filter	ImageNet	0.242	0.292
OpenAI ViT-B/32 Image Binary Filter	ImageNet	0.266	0.295
ResNet-34 Image Binary Filter	CC12M	0.203	0.257
OpenAI ViT-B/32 Image Binary Filter	CC12M	0.218	0.276
M3AE ViT-B/16	CC12M	0.237	0.297
CLIP ViT-B/32	CC12M	0.289	0.335

Таблица 1

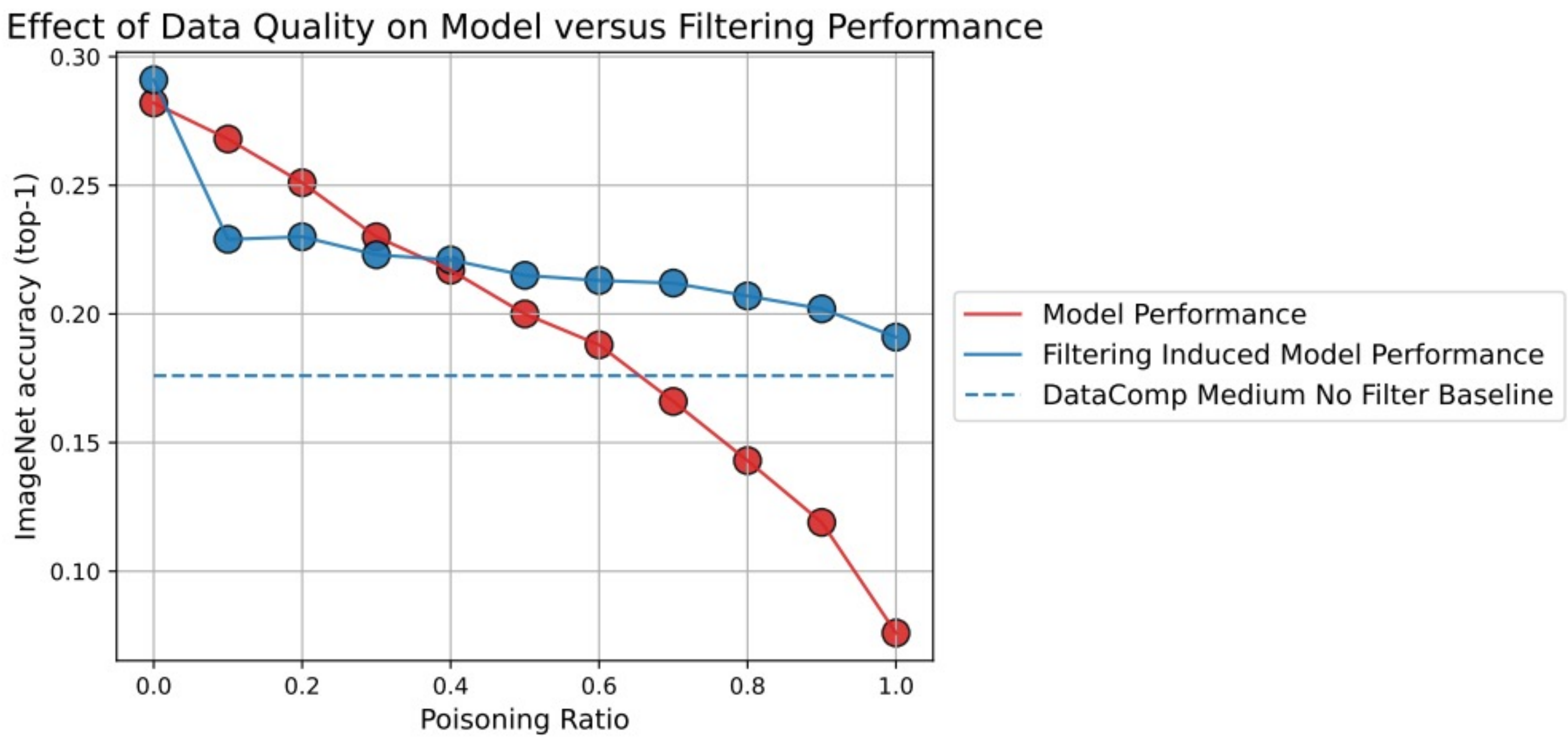


Рис. 2

DFNs: Data Filtering Networks (Fang A. et al. 2023)

Итоговый пайплайн:

1. **Обучаем DFN (ViT-B/32 CLIP) на High-Quality Image-Text Pairs (HQIMTP-350M).**
2. **После этого делают finetune на MS COCO, Flickr30k, ImageNet1k.**
3. **С помощью обученной DFN оставляют топ-15% пар из Common Crawl.**

Также авторы показывают, что обучение на CC12M + CC3M + SS15M тоже дает хороший результат.

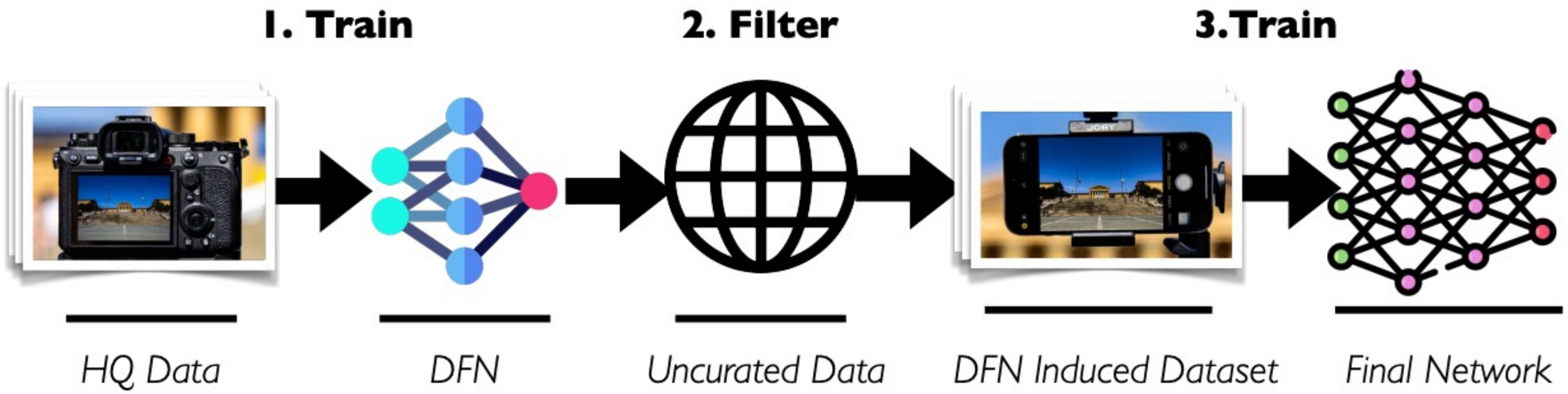


Рис. 4

BYOD track. Medium

Rank	Created	Submission	ImageNet acc.	Average perf.
1	09-06-2023	Image-cluster and CLIP (40%) + CC12M (50%) + Eval_trainsets (MNIST*3)	0.326	0.398
2	09-06-2023	CLIP (30%) + CC12M (50%) + Eval_trainsets (MNIST*3)	0.285	0.390
3	08-25-2023	Image-based intersect (CLIP score (L/14 30%) and BLIP2 (L/14 75%))	0.347	0.375
4	08-03-2023	CLIP score (L/14 30%) and BLIP2 (remaining 70%, filtered)	0.318	0.373
5	04-28-2023	Baseline: 4 external sources	0.36	0.345
6	04-28-2023	Baseline: Shutterstock 15M	0.229	0.29
7	04-28-2023	Baseline: CC12M	0.245	0.272
8	04-28-2023	Baseline: RedCaps	0.237	0.263
9	04-28-2023	Baseline: YFCC15M	0.232	0.257

Improving multimodal datasets with image captioning (Nguyen T. et al. 2023)

BYOD track. Top-1 L. Top-3 M.

Генерируют новые подписи к картинкам.

Если для генерации брать модель, обученную под метрики качества генерации подписей, метрики качества модели, обученной на сгенерированных таким образом подписях будут хуже (Таблица 2).

Использование сгенерированных BLIP2 подписей дает лучшее распределение CLIP score (Рис. 5).

Captioning model	NoCaps CIDEr [51]	CLIP-S [21]	Cosine similarity	No. of unique trigrams	ImageNet accuracy	Flickr retrieval
BLIP, ViT-L/16 (finetuned)	113.2*	0.698	0.231	2.82×10^6	0.207	0.498
BLIP2, ViT-g	80.6	0.737	0.251	2.72×10^6	0.281	0.507
BLIP2, ViT-g (finetuned)	119.7*	0.711	0.235	1.97×10^6	0.227	**0.549**
OpenCLIP-CoCa, ViT-L/14	0.354*	0.752	0.260	4.45×10^6	**0.321**	0.395
OpenCLIP-CoCa, ViT-L/14 (finetuned)	106.5*	0.702	0.232	1.81×10^6	0.252	0.542

Таблица 2

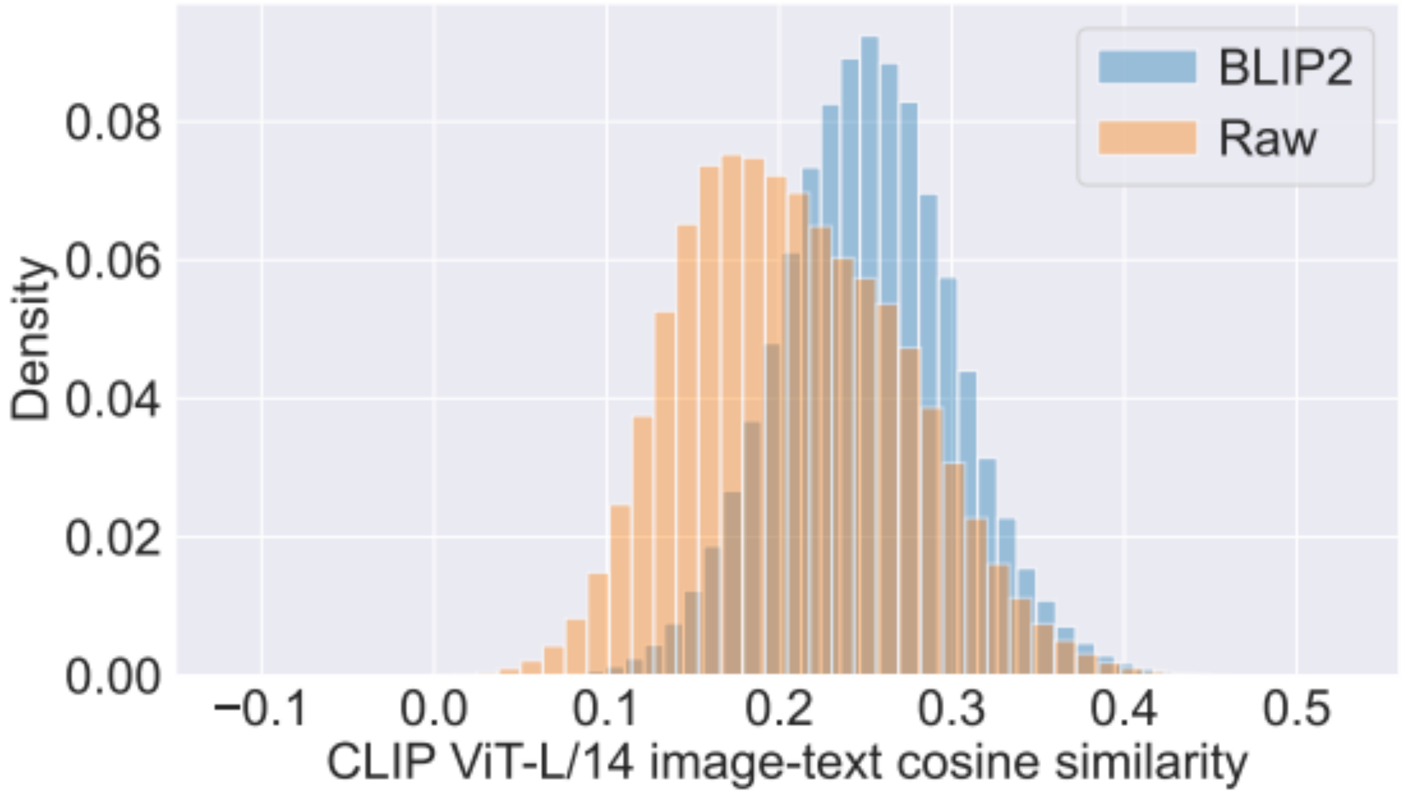


Рис. 5

Improving multimodal datasets with image captioning (Nguyen T. et al. 2023)

Images – original captions

Common Craw data. Оцениваем с помощью CLIP score + оцениваем по ImageNet1k clustering.

Images – good original captions (30%)

Images – synthetic captions (70%)

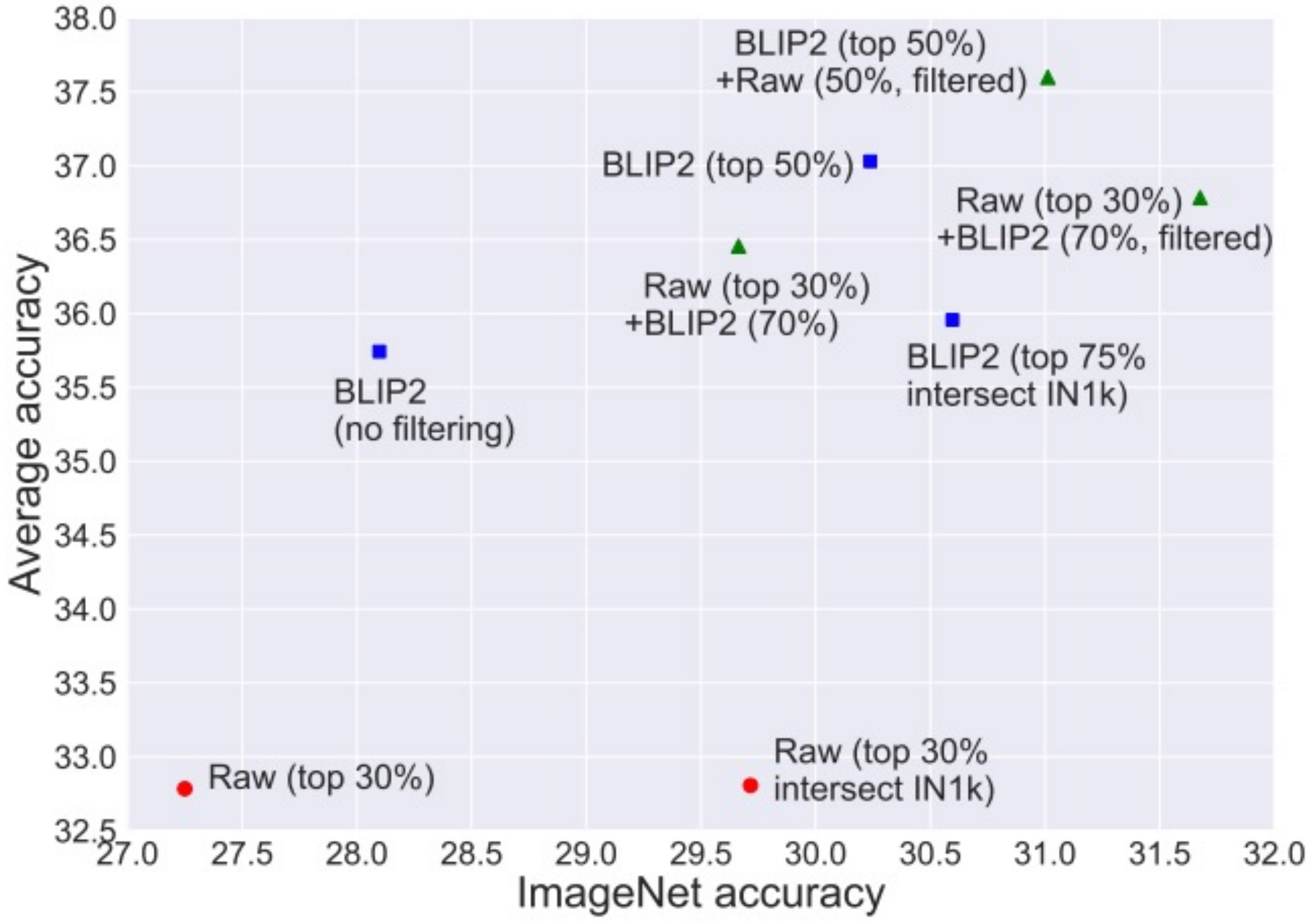
Генерируем подписи BLIP2

Images – synthetic captions (70%)

Оцениваем CLIP score

Images – good synthetic captions

Images – bad synthetic captions



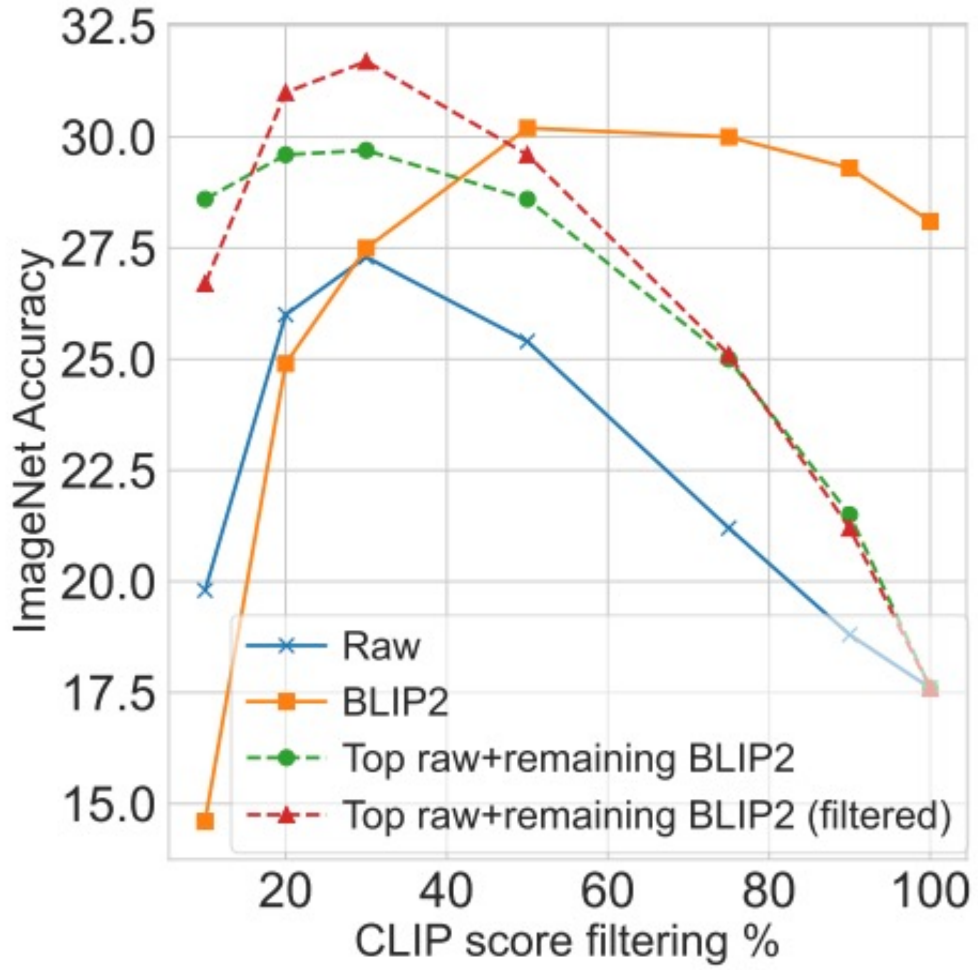


Рис. 6

SIEVE: Multimodal Dataset Pruning Using Image Captioning Models (Mahmoud A. et al. 2023)

Filtering track. Top-2 L. Top-8 M.

Генерируют новые подписи к картинкам (BLIP). Оценивают сходство сгенерированных подписей с оригинальными (all-MiniLM-L6-v2). По этой оценке фильтруют данные.

$$f_{\text{SIEVE}}(I,T) = \max_{T_j^G \in G(I,r)} \langle S(M(T_j^G)), S(M(T)) \rangle$$

$$f_{\text{SIEVE+CLIP}}(I,T) = (1-\alpha) \times \overline{f}_{\text{SIEVE}}(I,T) + \alpha \times \overline{f}_{\text{CLIP}}(I,T)$$

Маскируют общие слова для подписей (‘image of’, ‘picture of’).

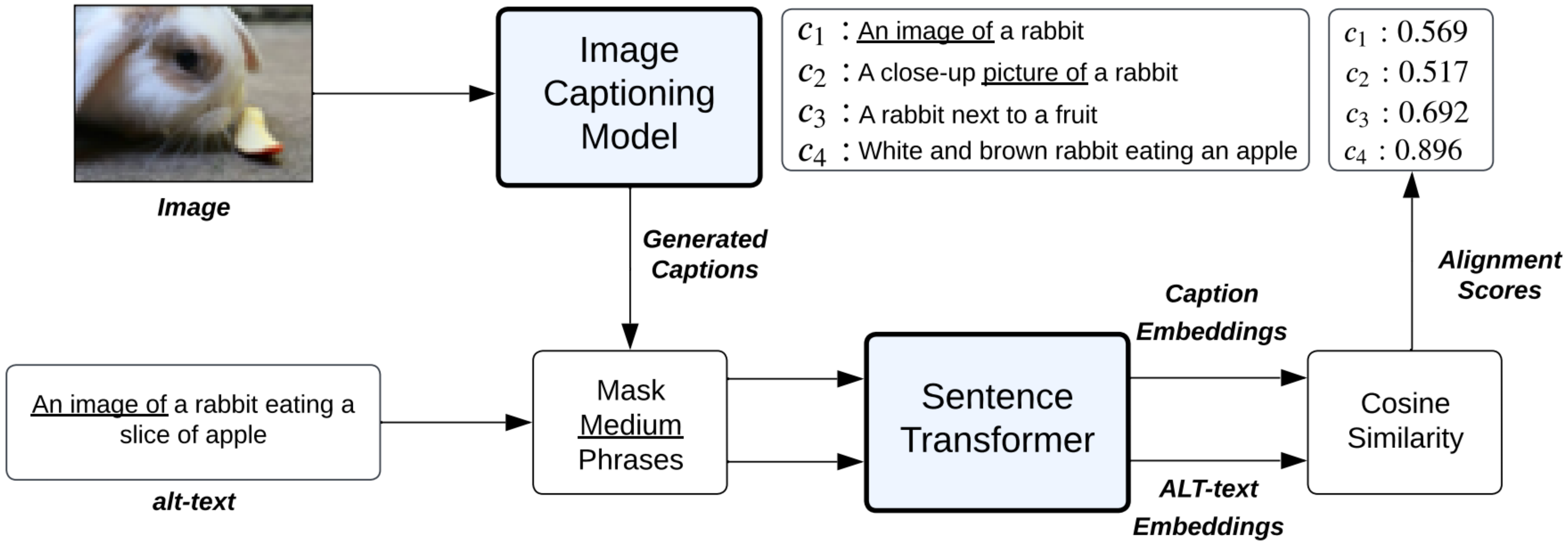


Рис. 7

Filtering track. Medium

Rank	Created	Submission	ImageNet acc.	Average perf.
1	11-08-2023	Hype sampler + DFN	0.382	0.379
2	11-07-2023	Hype sampler	0.346	0.373
3	10-02-2023	Data Filtering Networks	0.371	0.373
4	09-08-2023	The Devil Is in the Details	0.320	0.371
5	09-08-2023	TMARS + SSFT	0.338	0.362
6	08-17-2023	T-MARS: Improving Visual Representations by Circumventing Text Feature Learning	0.330	0.361
7	09-08-2023	The Devil Is in the Details - ImageNet best	0.336	0.355
8	08-25-2023	SIEVE	0.303	0.354
9	09-05-2023	Density-based Self-supervised Prototypes Pruning	0.334	0.345
10	09-07-2023	OCR and Naive english filtering	0.294	0.343
11	08-22-2023	WS (baselines)	0.305	0.342
12	07-26-2023	Mixed rules	0.303	0.337
13	04-28-2023	Baseline: Image-based ∩ CLIP score (L/14 30%)	0.297	0.328

Filtering track. Large, ExtraLarge

Rank	Created	Submission	ImageNet acc.	Average perf.
1	10-02-2023	Data Filtering Networks	0.678	0.560
2	08-25-2023	SIEVE	0.597	0.546
3	04-28-2023	Baseline: Image-based ∩ CLIP score (L/14 30%)	0.631	0.537
4	04-28-2023	Baseline: CLIP score (L/14 30%)	0.578	0.529

Rank	Created	Submission	ImageNet acc.	Average perf.
1	10-02-2023	Data Filtering Networks	0.814	0.669
2	04-28-2023	Baseline: Image-based ∩ CLIP score (L/14 30%)	0.792	0.663
3	04-28-2023	Baseline: CLIP score (L/14 30%)	0.764	0.65

BYOD track. Medium

Rank	Created	Submission	ImageNet acc.	Average perf.
1	09-06-2023	Image-cluster and CLIP (40%) + CC12M (50%) + Eval_trainsets (MNIST*3)	0.326	0.398
2	09-06-2023	CLIP (30%) + CC12M (50%) + Eval_trainsets (MNIST*3)	0.285	0.390
3	08-25-2023	Image-based intersect (CLIP score (L/14 30%) and BLIP2 (L/14 75%))	0.347	0.375
4	08-03-2023	CLIP score (L/14 30%) and BLIP2 (remaining 70%, filtered)	0.318	0.373
5	04-28-2023	Baseline: 4 external sources	0.36	0.345
6	04-28-2023	Baseline: Shutterstock 15M	0.229	0.29
7	04-28-2023	Baseline: CC12M	0.245	0.272
8	04-28-2023	Baseline: RedCaps	0.237	0.263
9	04-28-2023	Baseline: YFCC15M	0.232	0.257

BYOD track. Large, ExtraLarge

Rank	Created	Submission	ImageNet acc.	Average perf.
1	08-03-2023	Image-based intersect (CLIP score (L/14 30%) and BLIP2 (remaining 70%, filtered))	0.643	0.549
2	04-28-2023	Baseline: CommonPool CLIP score filter + 4 external sources (upsampled 2x)	0.621	0.541
3	04-28-2023	Baseline: CommonPool CLIP score filter + 4 external sources	0.609	0.536
4	04-28-2023	Baseline: LAION-2B	0.585	0.515

Rank	Created	Submission	ImageNet acc.	Average perf.
1	04-28-2023	Baseline: CommonPool CLIP score filter + 4 external sources (upsampled 6x)	0.776	0.649
2	04-28-2023	Baseline: LAION-2B	0.757	0.621

Список литературы и источников

- Yu H. et al. The Devil is in the Details: A Deep Dive into the Rabbit Hole of Data Filtering //arXiv preprint arXiv:2309.15954. – 2023.
- Fang A. et al. Data Filtering Networks //arXiv preprint arXiv:2309.17425. – 2023.
- Nguyen T. et al. Improving multimodal datasets with image captioning //arXiv preprint arXiv:2307.10350. – 2023.
- Mahmoud A. et al. SIEVE: Multimodal Dataset Pruning Using Image Captioning Models //arXiv preprint arXiv:2310.02110. – 2023.