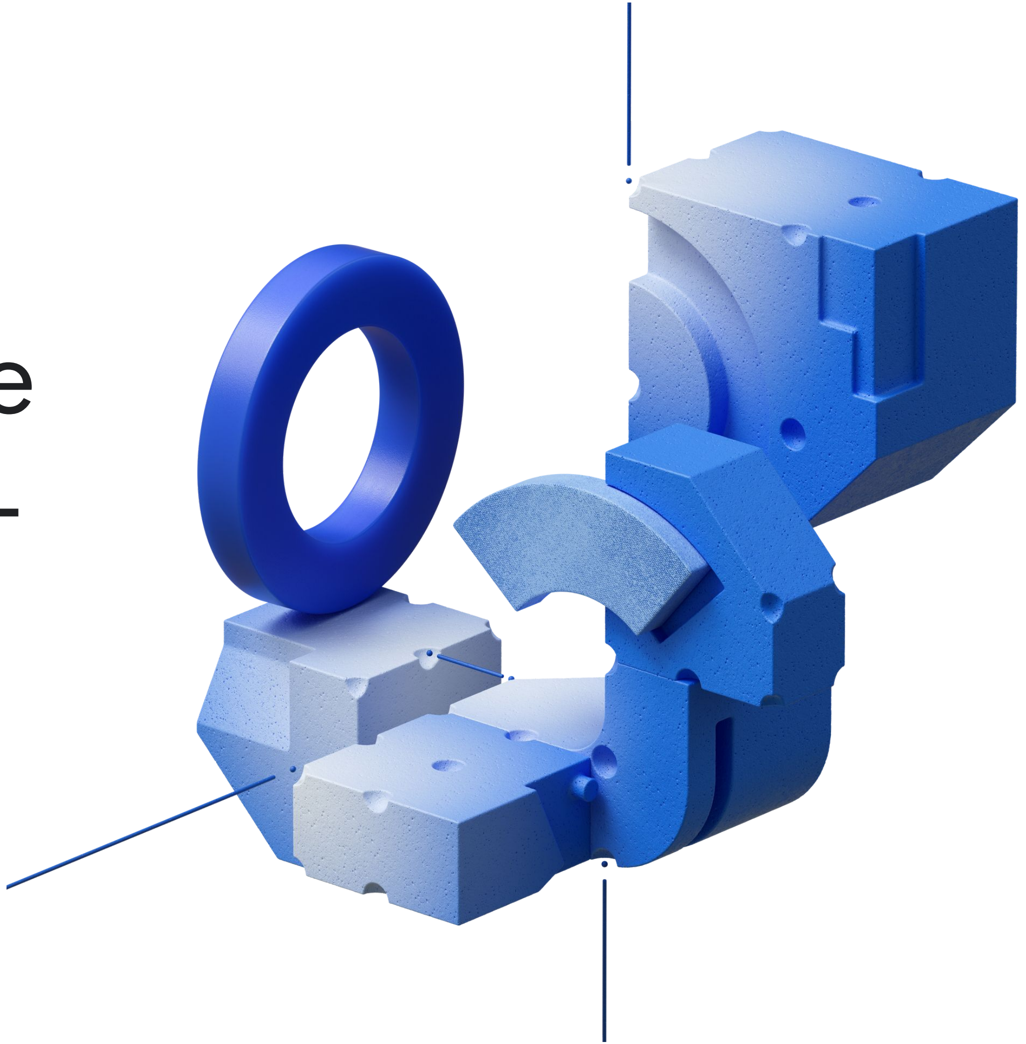
Google DeepMind

Data curation is the next frontier of SSL

Olivier Hénaff

ECCV 2024 SSL WIN Workshop

29 September 2024



The world runs on self-supervised learning

Unimodal SSL on

- images (e.g. DINOv2)
- text (e.g. GPT-3)

Multimodal SSL

- image-text (e.g. CLIP)
- video-audio (e.g. MMV)

The world runs on self-supervised learning

Unimodal SSL on

- images (e.g. DINOv2)
- text (e.g. GPT-3)

Multimodal SSL

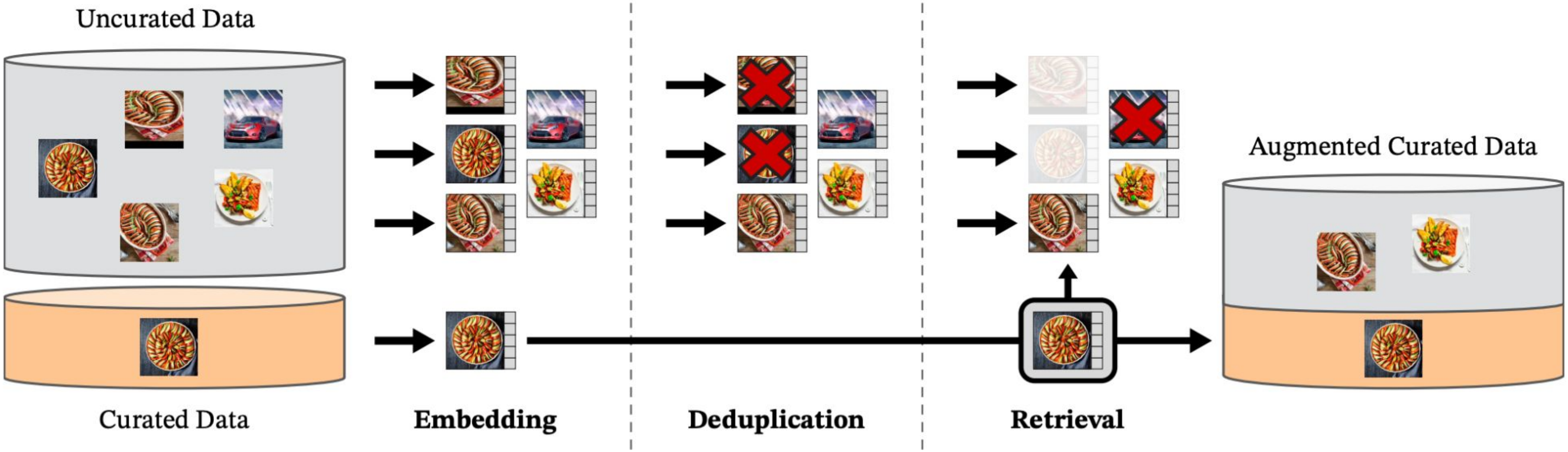
- image-text (e.g. CLIP)
- video-audio (e.g. MMV)

... ***all rests on clever choices of data***

The world runs on self-supervised learning + data curation

Oquab, 2023

Image SSL with DINOv2: strong curation with eval data

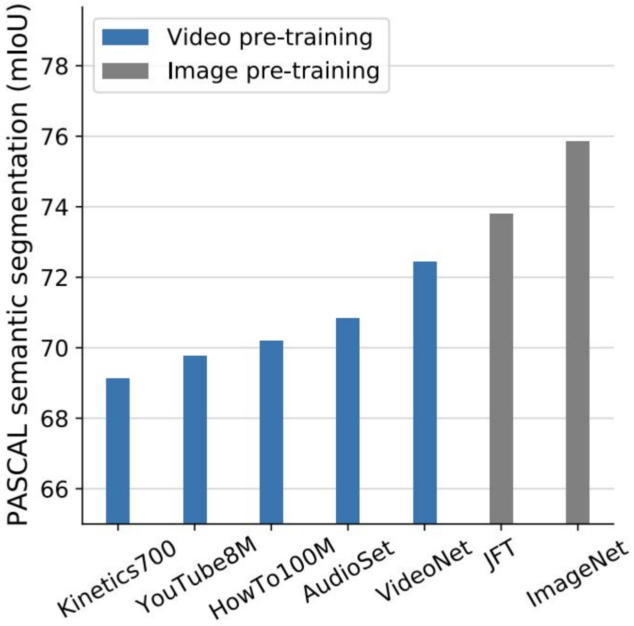


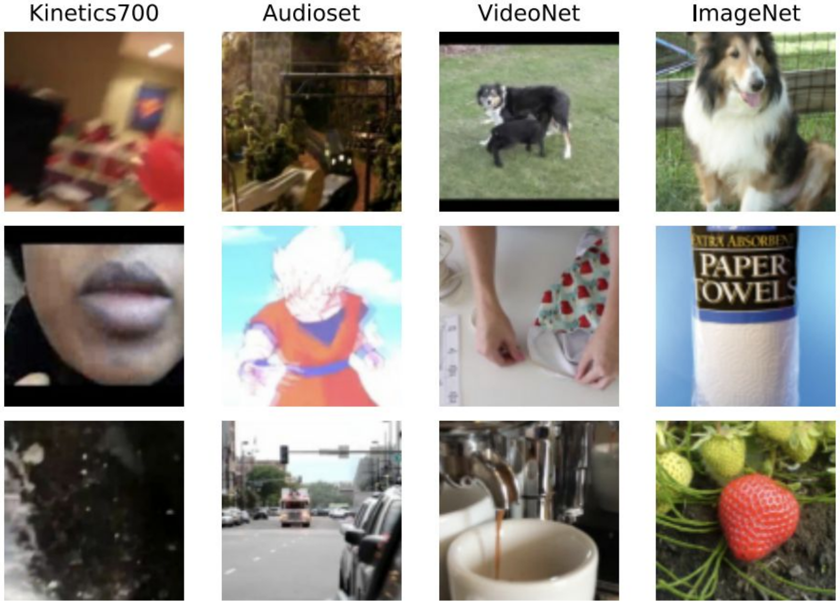
Google

The world runs on self-supervised learning + data curation

Parthasarathy, 2023

Video SSL with VITO: curation with high-quality image prior

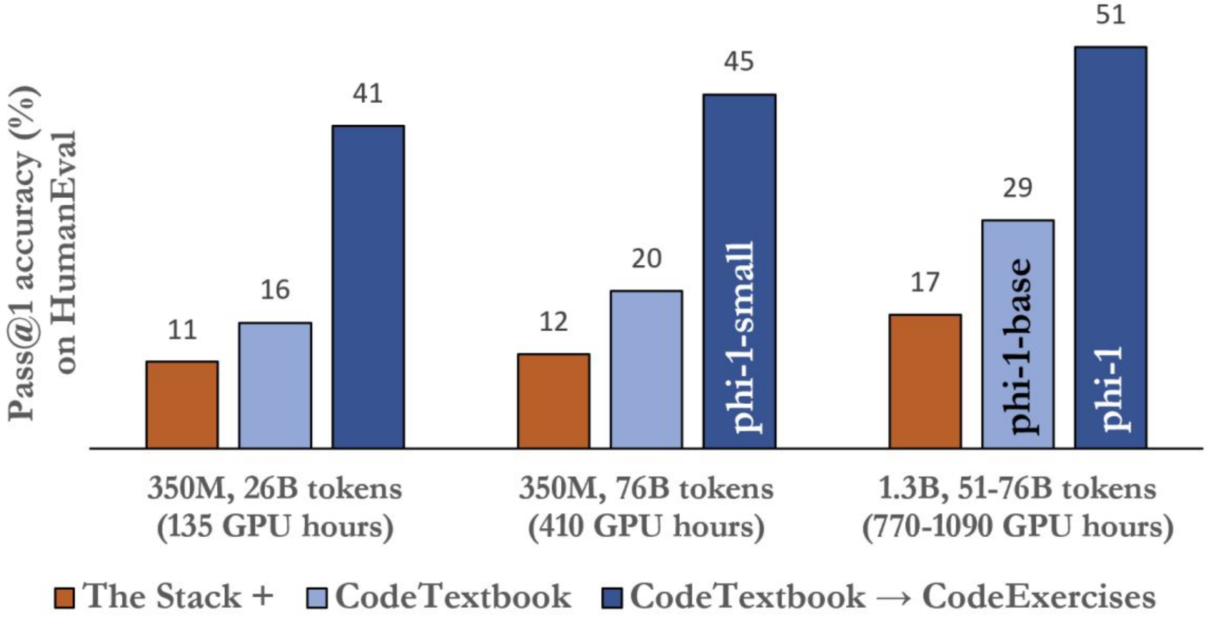




The world runs on self-supervised learning + data curation

Gunasekar, 2023

Current LLM's are highly dependent on data quality



The world runs on self-supervised learning + data curation

Dataset	Pretraining (as is)	Retrieving pretraining data	Eval.	Task	Citation
ImageNet-1k	✗	✓	✓	Classif.	(Russakovsky et al., 2015)
ImageNet-22k	✓	✓	✗		(Deng et al., 2009)
ImageNet-V2	✗	✗	✓	Classif.	(Recht et al., 2019)
ImageNet-ReaL	✗	✗	✓	Classif.	(Beyer et al., 2020)
ImageNet-A	✗	✗	✓	Classif.	(Hendrycks et al., 2021b)
ImageNet-C	✗	✗	✓	Classif.	(Hendrycks & Dietterich, 2019)
ImageNet-R	✗	✗	✓	Classif.	(Hendrycks et al., 2021a)
ImageNet-Sk.	✗	✗	✓	Classif.	(Wang et al., 2019)
Food-101	✗	✓	✓	Classif.	(Bossard et al., 2014)
CIFAR-10	✗	✓	✓	Classif.	(Krizhevsky et al., 2009)
CIFAR-100	✗	✓	✓	Classif.	(Krizhevsky et al., 2009)
SUN397	✗	✓	✓	Classif.	(Xiao et al., 2010)
StanfordCars	✗	✓	✓	Classif.	(Krause et al., 2013)
FGVC-Aircraft	✗	✓	✓	Classif.	(Maji et al., 2013)
VOC 2007	✗	✓	✓	Classif.	(Everingham et al., 2010)
DTD	✗	✓	✓	Classif.	(Cimpoi et al., 2014)
Oxford Pets	✗	✓	✓	Classif.	(Parkhi et al., 2012)
Caltech101	✗	✓	✓	Classif.	(Fei-Fei et al., 2004)
Flowers	✗	✓	✓	Classif.	(Nilsback & Zisserman, 2008)
CUB200	✗	✓	✓	Classif.	(Welinder et al., 2010)
iNaturalist 2018	✗	✗	✓	Classif.	(Van Horn et al., 2018)
iNaturalist 2021	✗	✗	✓	Classif.	(Van Horn et al., 2021)
Places-205	✗	✗	✓	Classif.	(Zhou et al., 2014)
UCF101	✗	✗	✓	Video	(Soomro et al., 2012)
Kinetics-400	✗	✗	✓	Video	(Kay et al., 2017)
SSv2	✗	✗	✓	Video	(Goyal et al., 2017)
GLD v2	✓	✓	✗		(Weyand et al., 2020)
R-Paris	✗	✓	✓	Retrieval	(Radenović et al., 2018a)
R-Oxford	✗	✓	✓	Retrieval	(Radenović et al., 2018a)
Met	✗	✓	✓	Retrieval	(Ypsilantis et al., 2021)
Amstertime	✗	✓	✓	Retrieval	(Yildiz et al., 2022)
ADE20k	✗	✓	✓	Seg.	(Zhou et al., 2017)
Cityscapes	✗	✓	✓	Seg.	(Cordts et al., 2016)
VOC 2012	✗	✓	✓	Seg.	(Everingham et al., 2010)
Mapillary SLS	✓	✗	✗		(Warburg et al., 2020)
NYU-Depth V2	✗	✓	✓	Depth	(Silberman et al., 2012)
KITTI	✗	✓	✓	Depth	(Geiger et al., 2013)
SUN-RGBD	✗	✓	✓	Depth	(Song et al., 2015)
DollarStreet	✗	✗	✓	Fairness	(De Vries et al., 2019)
Casual Conv.	✗	✗	✓	Fairness	(Hazirbas et al., 2021)

Yet data-curation is currently a secretive & tedious process

- More “feature engineering” than “deep learning”
- Lots of details hidden in appendices
- Hard to reproduce specific dataset versions

The world runs on self-supervised learning + data curation

Dataset	Pretraining (as is)	Retrieving pretraining data	Eval.	Task	Citation
ImageNet-1k	✗	✓	✓	Classif.	(Russakovsky et al., 2015)
ImageNet-22k	✓	✓	✗		(Deng et al., 2009)
ImageNet-V2	✗	✗	✓	Classif.	(Recht et al., 2019)
ImageNet-ReaL	✗	✗	✓	Classif.	(Beyer et al., 2020)
ImageNet-A	✗	✗	✓	Classif.	(Hendrycks et al., 2021b)
ImageNet-C	✗	✗	✓	Classif.	(Hendrycks & Dietterich, 2019)
ImageNet-R	✗	✗	✓	Classif.	(Hendrycks et al., 2021a)
ImageNet-Sk.	✗	✗	✓	Classif.	(Wang et al., 2019)
Food-101	✗	✓	✓	Classif.	(Bossard et al., 2014)
CIFAR-10	✗	✓	✓	Classif.	(Krizhevsky et al., 2009)
CIFAR-100	✗	✓	✓	Classif.	(Krizhevsky et al., 2009)
SUN397	✗	✓	✓	Classif.	(Xiao et al., 2010)
StanfordCars	✗	✓	✓	Classif.	(Krause et al., 2013)
FGVC-Aircraft	✗	✓	✓	Classif.	(Maji et al., 2013)
VOC 2007	✗	✓	✓	Classif.	(Everingham et al., 2010)
DTD	✗	✓	✓	Classif.	(Cimpoi et al., 2014)
Oxford Pets	✗	✓	✓	Classif.	(Parkhi et al., 2012)
Caltech101	✗	✓	✓	Classif.	(Fei-Fei et al., 2004)
Flowers	✗	✓	✓	Classif.	(Nilsback & Zisserman, 2008)
CUB200	✗	✓	✓	Classif.	(Welinder et al., 2010)
iNaturalist 2018	✗	✗	✓	Classif.	(Van Horn et al., 2018)
iNaturalist 2021	✗	✗	✓	Classif.	(Van Horn et al., 2021)
Places-205	✗	✗	✓	Classif.	(Zhou et al., 2014)
UCF101	✗	✗	✓	Video	(Soomro et al., 2012)
Kinetics-400	✗	✗	✓	Video	(Kay et al., 2017)
SSv2	✗	✗	✓	Video	(Goyal et al., 2017)
GLD v2	✓	✓	✗		(Weyand et al., 2020)
R-Paris	✗	✓	✓	Retrieval	(Radenović et al., 2018a)
R-Oxford	✗	✓	✓	Retrieval	(Radenović et al., 2018a)
Met	✗	✓	✓	Retrieval	(Ypsilantis et al., 2021)
Amstertime	✗	✓	✓	Retrieval	(Yildiz et al., 2022)
ADE20k	✗	✓	✓	Seg.	(Zhou et al., 2017)
Cityscapes	✗	✓	✓	Seg.	(Cordts et al., 2016)
VOC 2012	✗	✓	✓	Seg.	(Everingham et al., 2010)
Mapillary SLS	✓	✗	✗		(Warburg et al., 2020)
NYU-Depth V2	✗	✓	✓	Depth	(Silberman et al., 2012)
KITTI	✗	✓	✓	Depth	(Geiger et al., 2013)
SUN-RGBD	✗	✓	✓	Depth	(Song et al., 2015)
DollarStreet	✗	✗	✓	Fairness	(De Vries et al., 2019)
Casual Conv.	✗	✗	✓	Fairness	(Hazirbas et al., 2021)

Yet data-curation is currently a secretive & tedious process

- More “feature engineering” than “deep learning”
- Lots of details hidden in appendices
- Hard to reproduce specific dataset versions

Let’s bring data curation to the front!

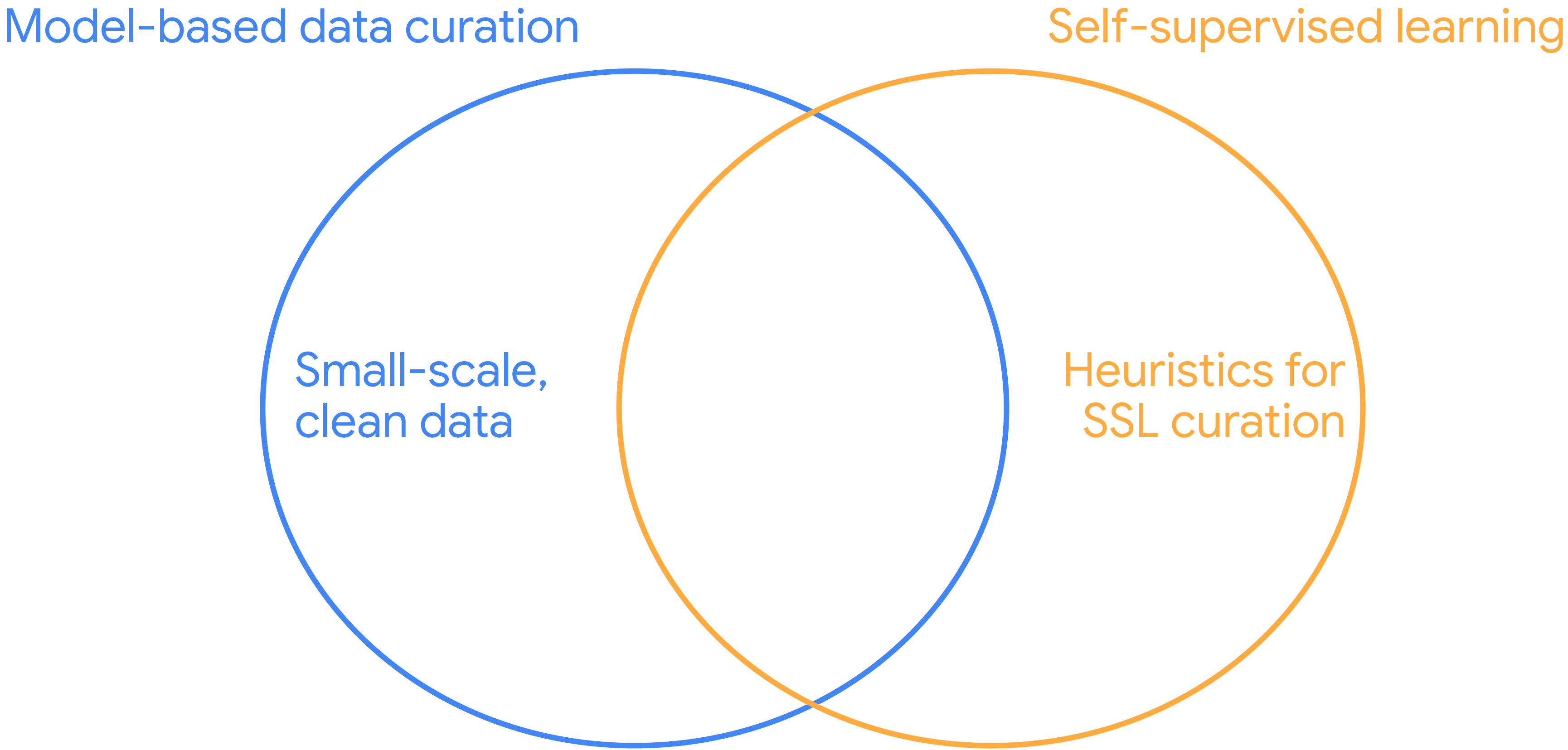
- Accept it as integral part of CV pipelines
- Own its details, allowing reproduction
- Same scientific rigor as architectures, objectives, optim

→ ***simple, scalable methods for data curation!***

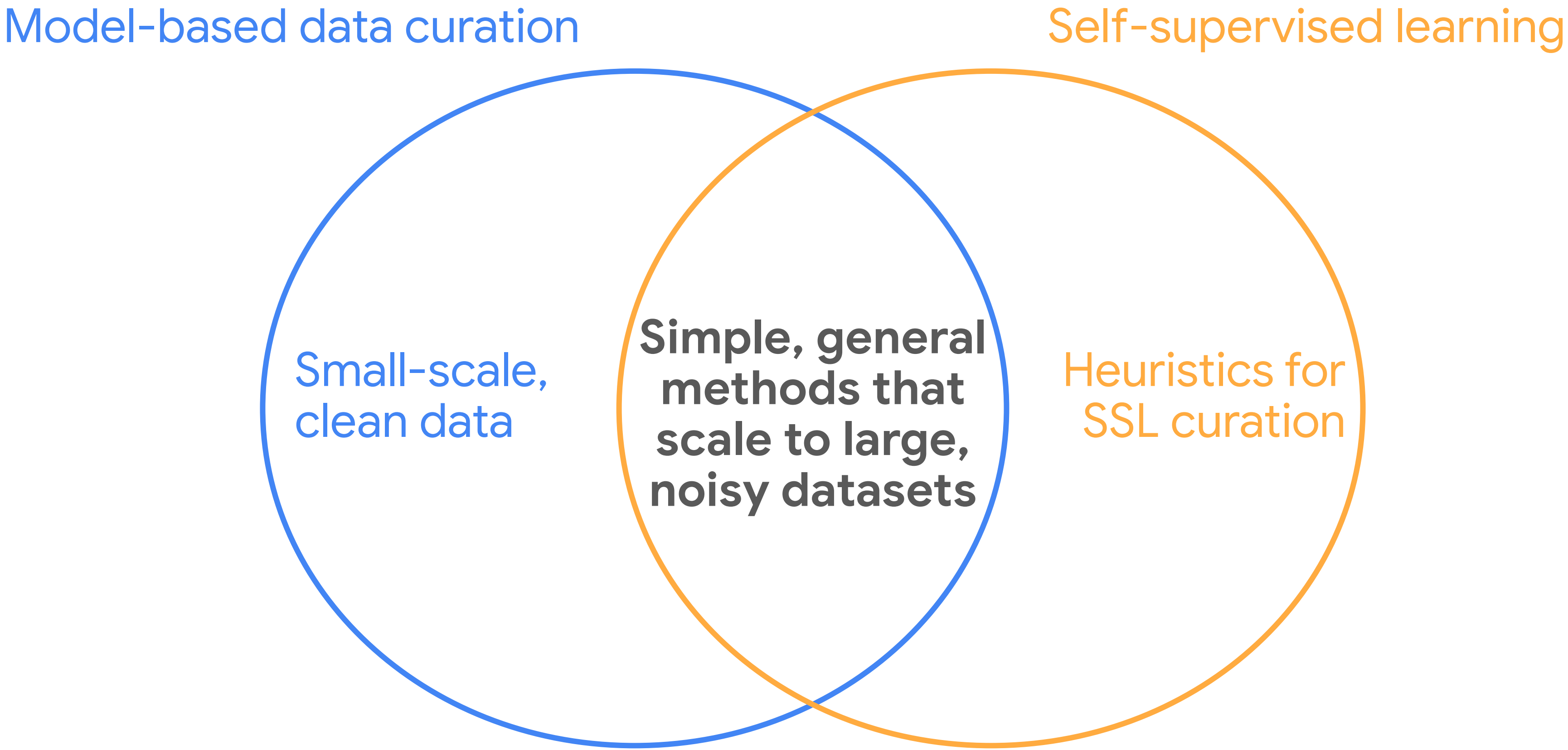
→ ***prime candidate: model-based data curation***

Google

Model-based data curation meets self-supervised learning



Model-based data curation meets self-supervised learning



Google

Model-based data curation meets self-supervised learning

Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding

⟶ builds a framework model-based data selection

- Which model-based criteria for data-selection?
- How to make data-selection tractable?

Model-based data curation meets self-supervised learning

Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding

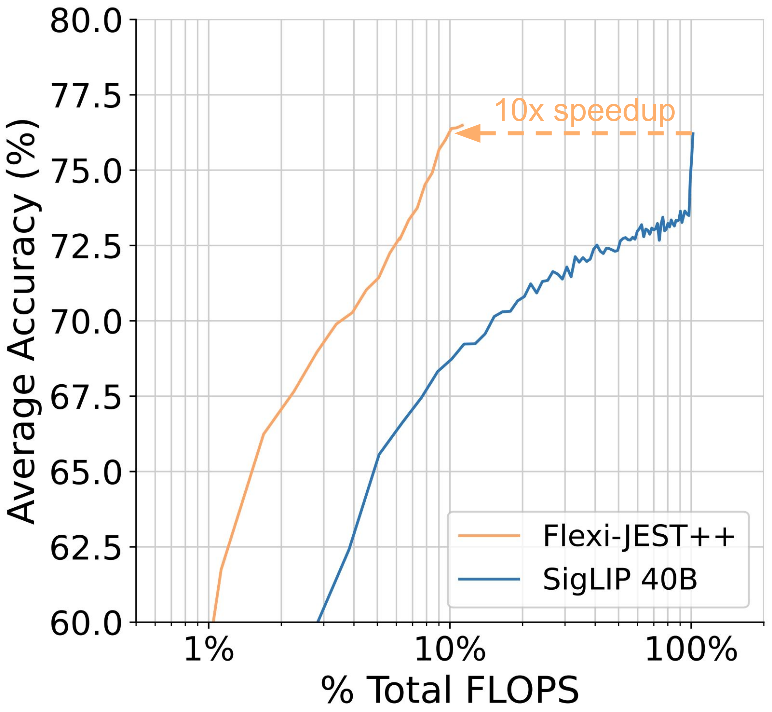
⟶ builds a framework model-based data selection

- Which model-based criteria for data-selection?
- How to make data-selection tractable?

Data Curation with Joint Example Selection Further Accelerates Multimodal Learning

⟶ applies this framework to multimodal contrastive SSL

- Contrastive SSL enables joint example selection (JEST)
- JEST radically accelerates multimodal learning (10x)

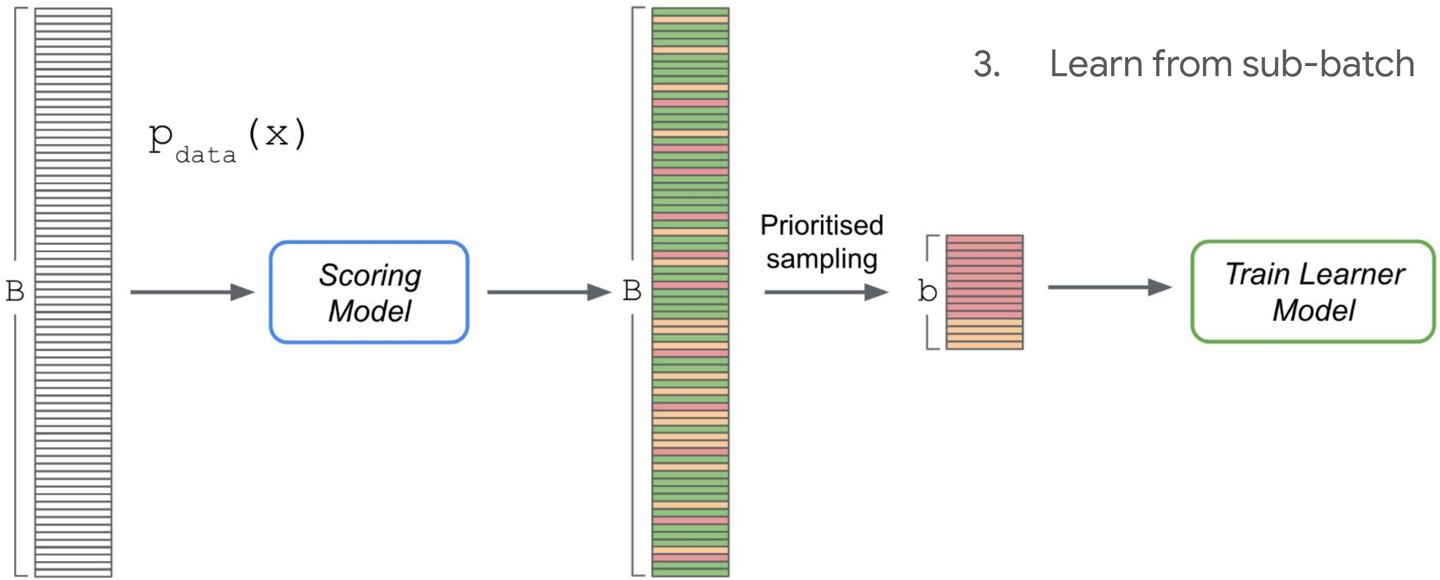


Google

Model-based data curation: framework

Data curation with online batch selection:

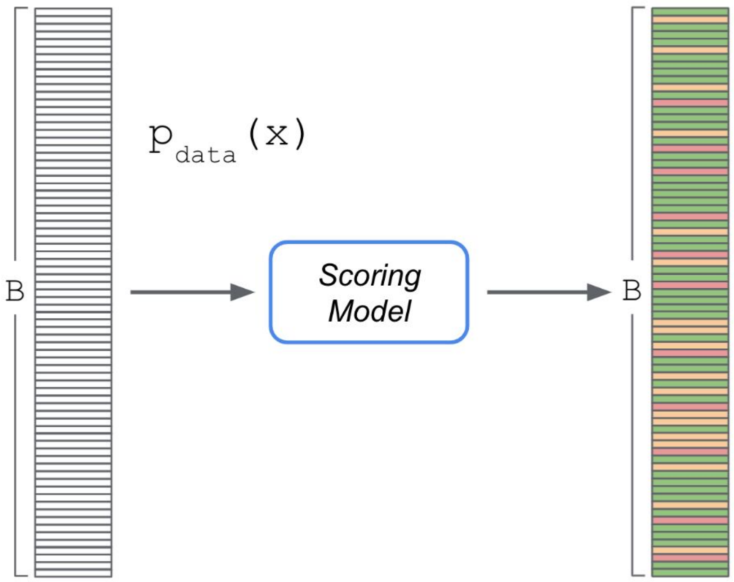
1. Score super-batch
2. Sub-sample batch according to these scores
3. Learn from sub-batch



Model-based data curation: criteria

Hard-learner: $s^{\text{hard}}(\boldsymbol{x}_i|\theta) = \ell(\boldsymbol{x}_i|\theta)$

→ removes trivial examples, but emphasizes noise



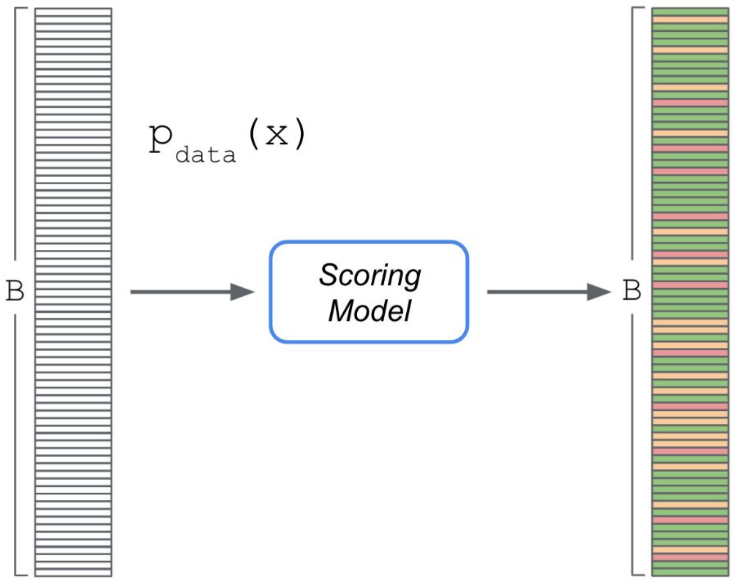
Model-based data curation: criteria

Hard-learner: $s^{\text{hard}}(\boldsymbol{x}_i|\theta) = \ell(\boldsymbol{x}_i|\theta)$

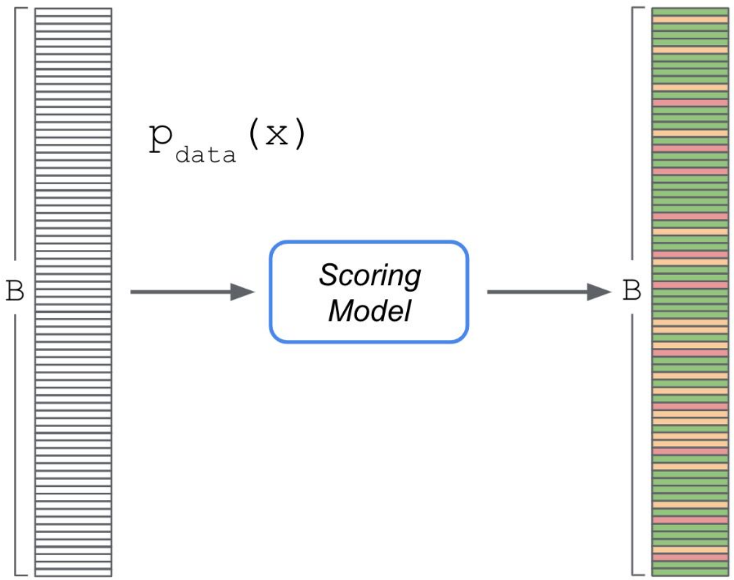
→ removes trivial examples, but emphasizes noise

Easy-reference: $s^{\text{easy}}(\boldsymbol{x}_i|\theta) = -\ell(\boldsymbol{x}_i|\theta)$ cf. CLIP-Score

→ removes noise, but emphasizes trivial examples



Model-based data curation: criteria



Hard-learner: $s^{\text{hard}}(\boldsymbol{x}_i|\theta) = \ell(\boldsymbol{x}_i|\theta)$

→ removes trivial examples, but emphasizes noise

Easy-reference: $s^{\text{easy}}(\boldsymbol{x}_i|\theta) = -\ell(\boldsymbol{x}_i|\theta)$ cf. CLIP-Score

→ removes noise, but emphasizes trivial examples

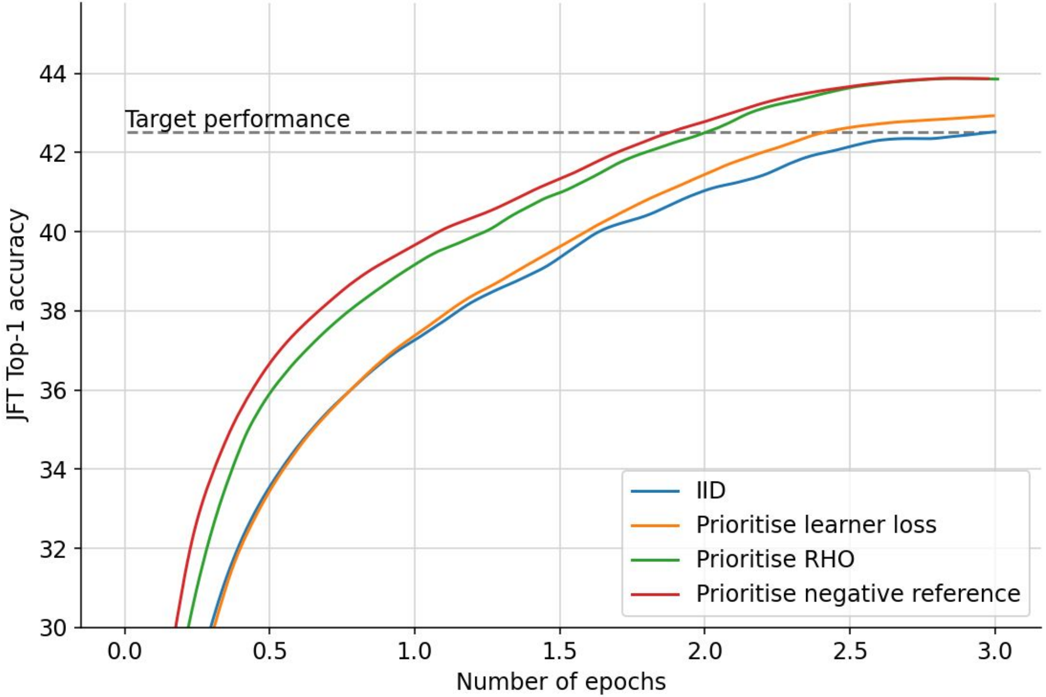
Learnability:
$$s^{\text{learn}}(\boldsymbol{x}_i|\theta^t, \theta^*) = s^{\text{hard}}(\boldsymbol{x}_i|\theta^t) + s^{\text{easy}}(\boldsymbol{x}_i|\theta^*)$$
$$= \ell(\boldsymbol{x}_i|\theta^t) - \ell(\boldsymbol{x}_i|\theta^*)$$

→ emphasizes hard examples that get easy with more compute (not trivial, not noisy)

Model-based data curation: criteria

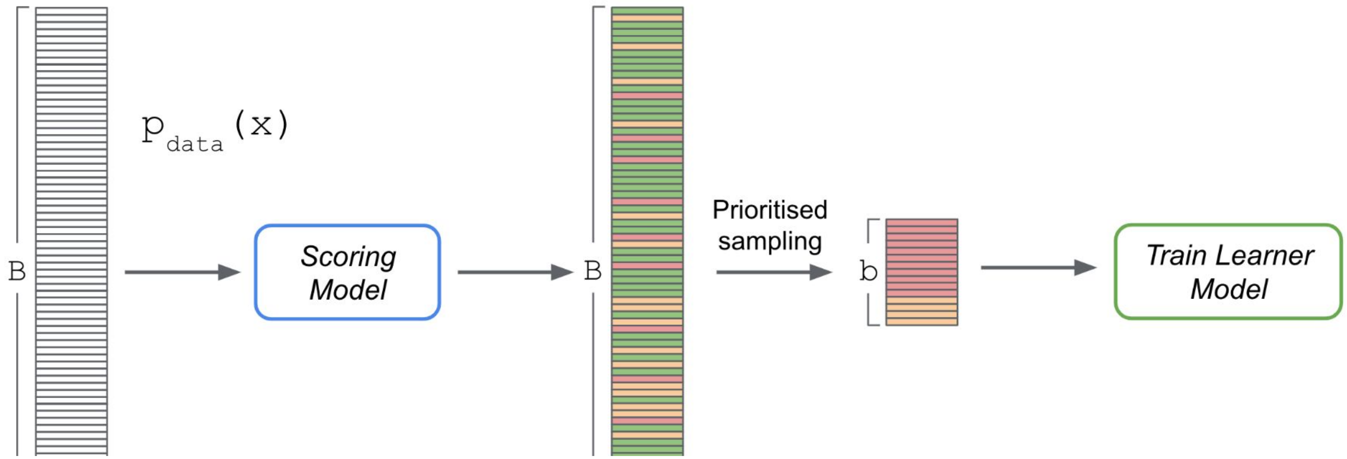
Large-scale classification on JFT-300M

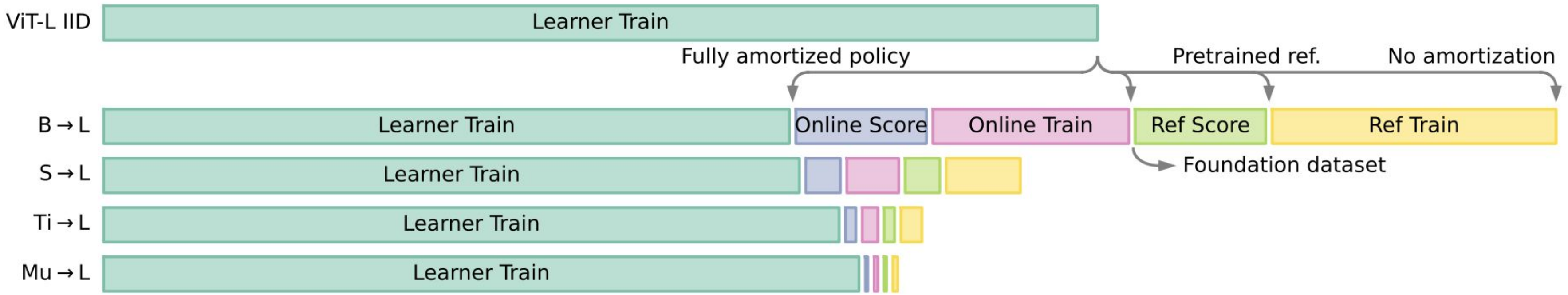
- Prioritize with **hard-learner**
  ⟶ 10% speed-up
- Prioritize **easy reference**
  ⟶ 30% speed-up
- Prioritize with **learnability**
  ⟶ 30% speed-up



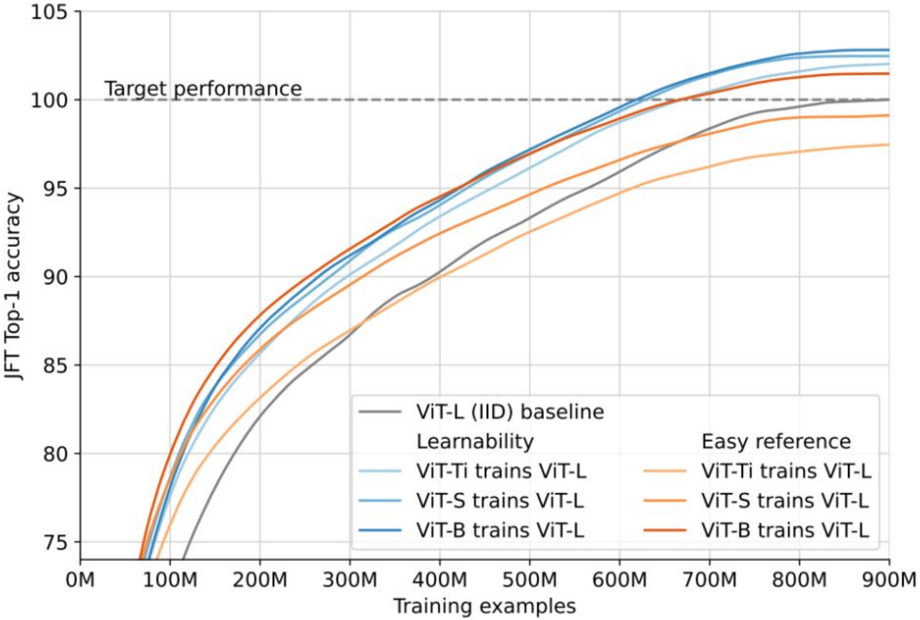
Model-based data curation: unlocking compute-positivity

Core intuition:
Curate data with small models, train big ones

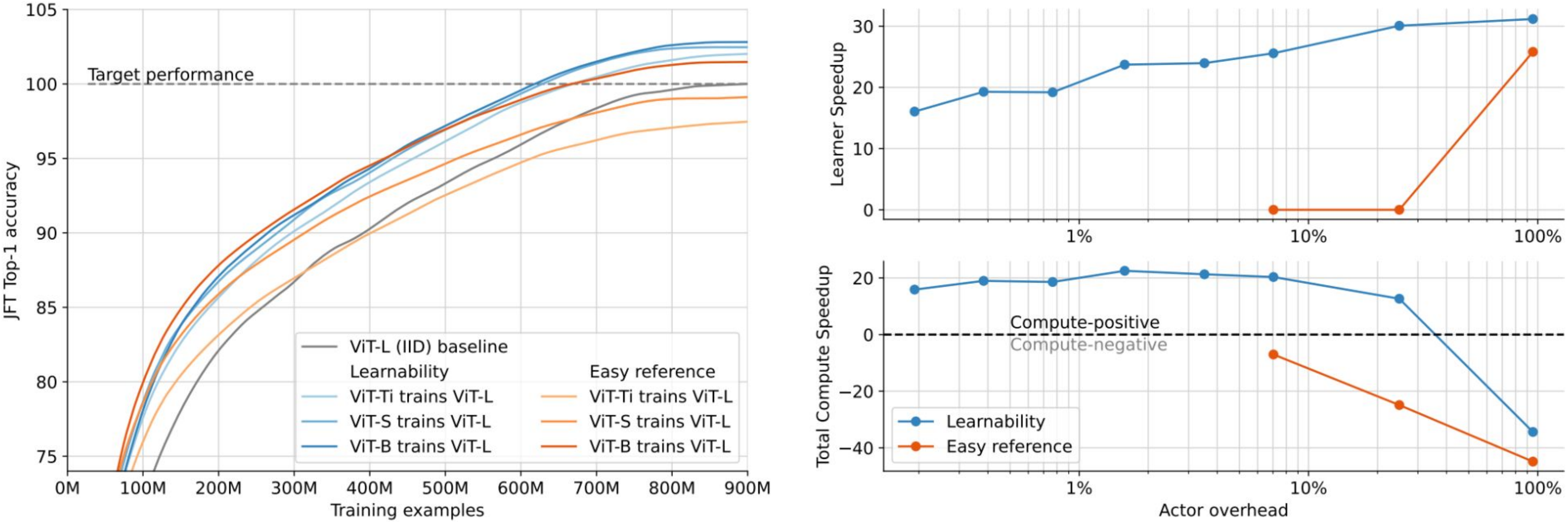




Model-based data curation: unlocking compute-positivity



Model-based data curation: unlocking compute-positivity



Model-based data curation meets self-supervised learning

Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding

⟶ builds a framework model-based data selection

- Which model-based criteria for data-selection? **⟶ learnability!**
- How to make data-selection tractable? **⟶ small models + generalizable policies!!**

Model-based data curation meets self-supervised learning

Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding

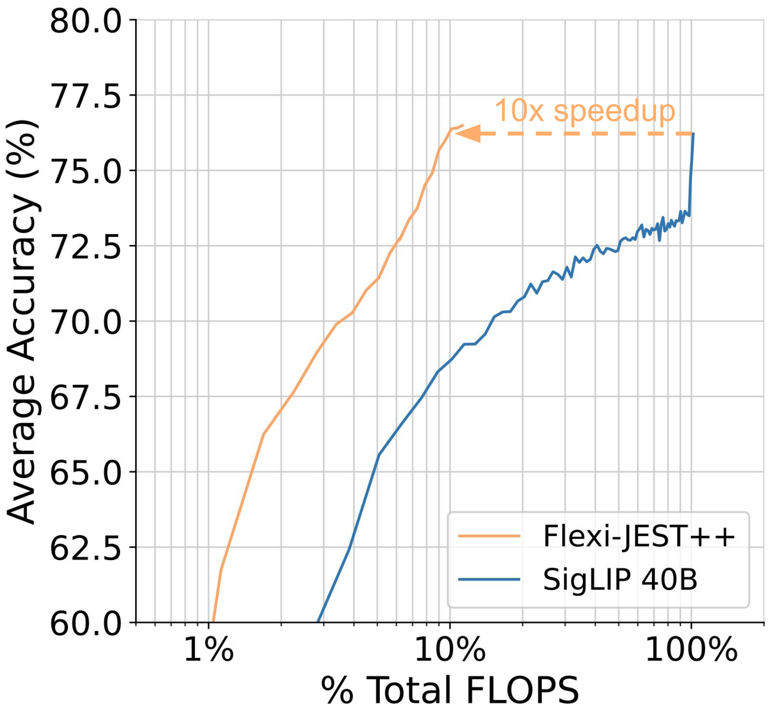
⟶ builds a framework model-based data selection

- Which model-based criteria for data-selection?
- How to make data-selection tractable?

Data Curation with Joint Example Selection Further Accelerates Multimodal Learning

⟶ applies this framework to multimodal contrastive SSL

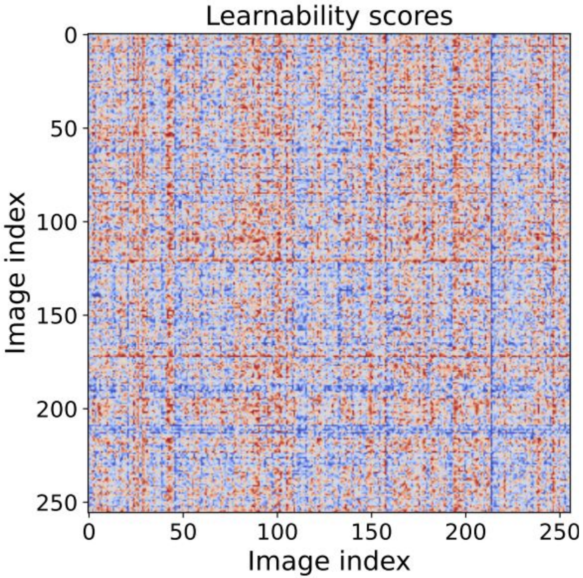
- Contrastive SSL enables joint example selection (JEST)
- JEST radically accelerates multimodal learning (10x)



Google

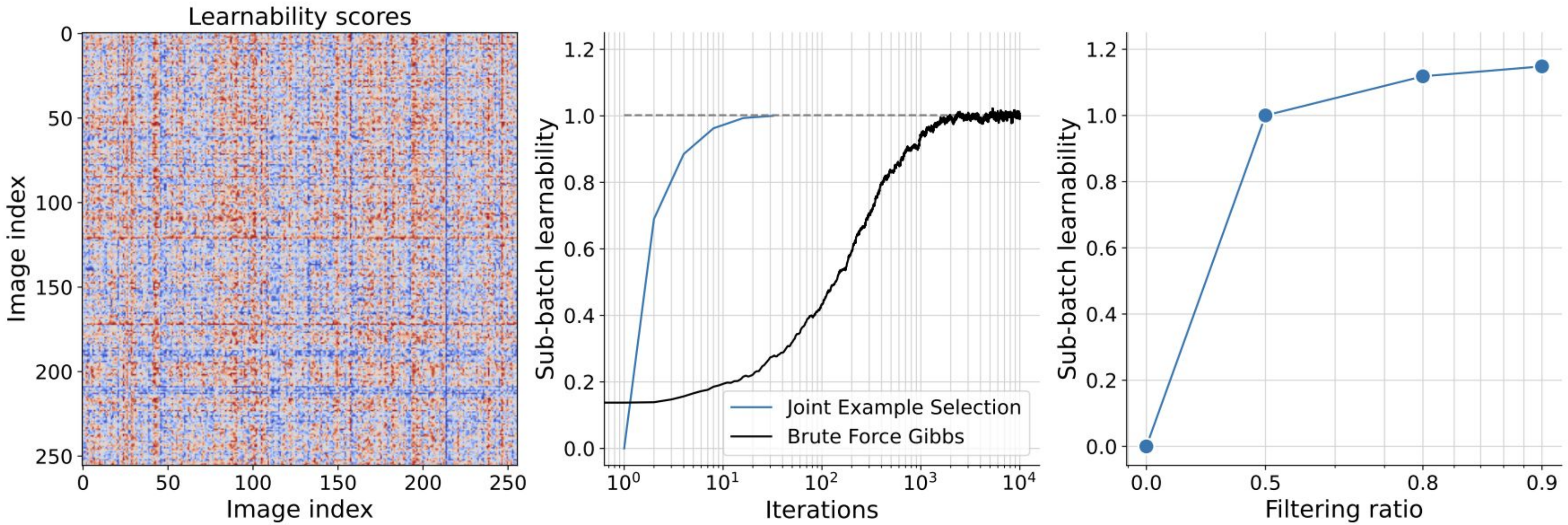
Joint Example Selection Accelerates Multimodal Learning

- **Model**: multimodal contrastive learning with SigLIP
- **Prior work**: only focuses on independent data selection, i.e. diagonals of the contrastive matrix
- **Intuition**: contrastive loss depends on entire matrix, and matrix is clearly non-diagonal!



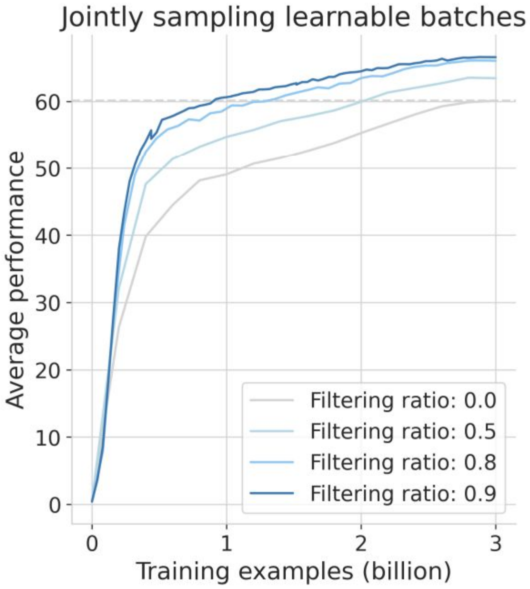
Joint Example Selection Accelerates Multimodal Learning

- **Model**: multimodal contrastive learning with SigLIP
- **Prior work**: only focuses on independent data selection, i.e. diagonals of the contrastive matrix
- **Intuition**: contrastive loss depends on entire matrix, and matrix is clearly non-diagonal!



Joint Example Selection Accelerates Multimodal Learning

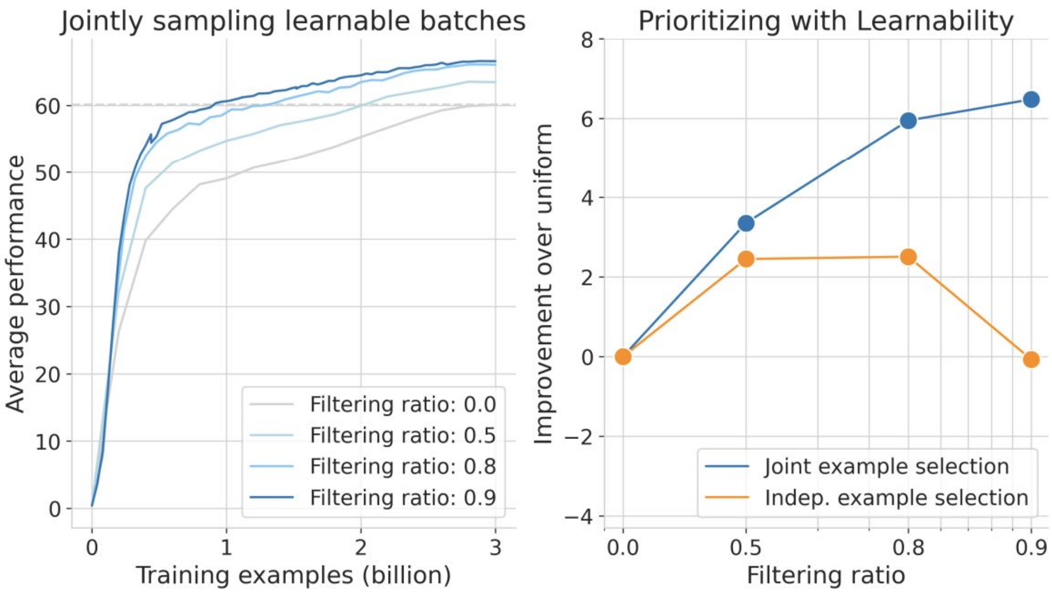
- **Model**: multimodal contrastive learning with SigLIP
- **Prior work**: only focuses on independent data selection, i.e. diagonals of the contrastive matrix
- **Intuition**: contrastive loss depends on entire matrix, and matrix is clearly non-diagonal!



Google

Joint Example Selection Accelerates Multimodal Learning

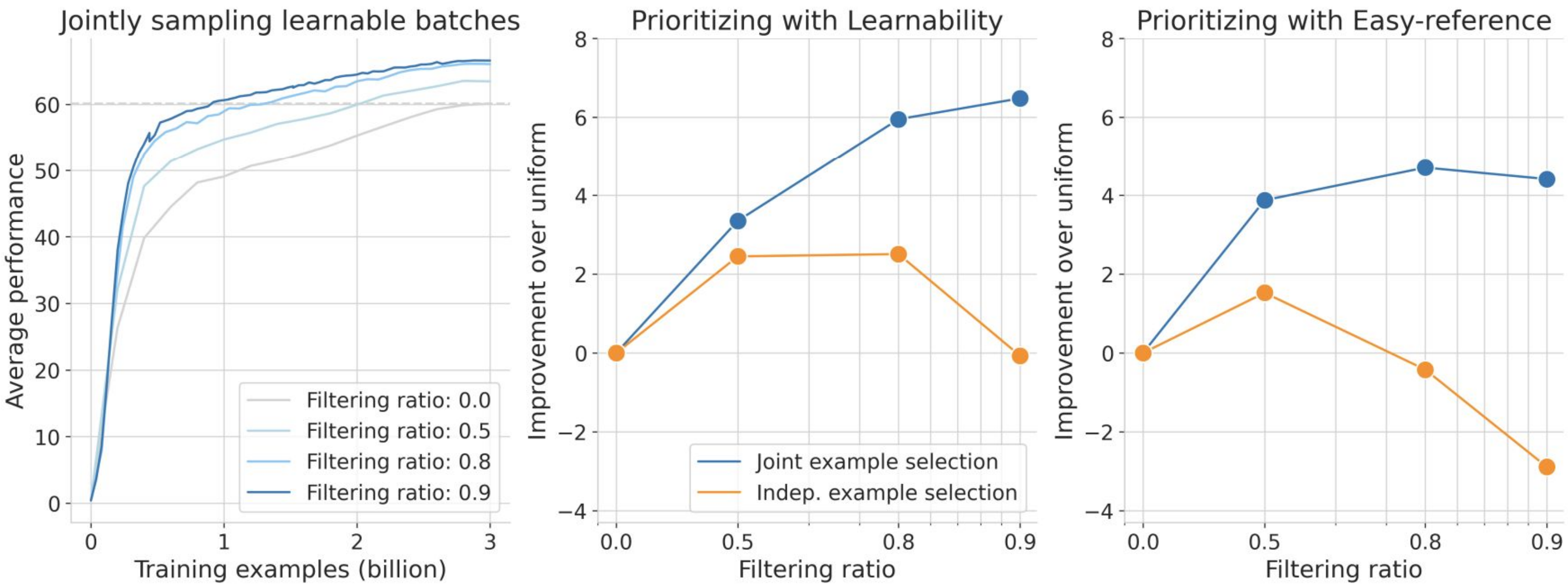
- **Model**: multimodal contrastive learning with SigLIP
- **Prior work**: only focuses on independent data selection, i.e. diagonals of the contrastive matrix
- **Intuition**: contrastive loss depends on entire matrix, and matrix is clearly non-diagonal!



Google

Joint Example Selection Accelerates Multimodal Learning

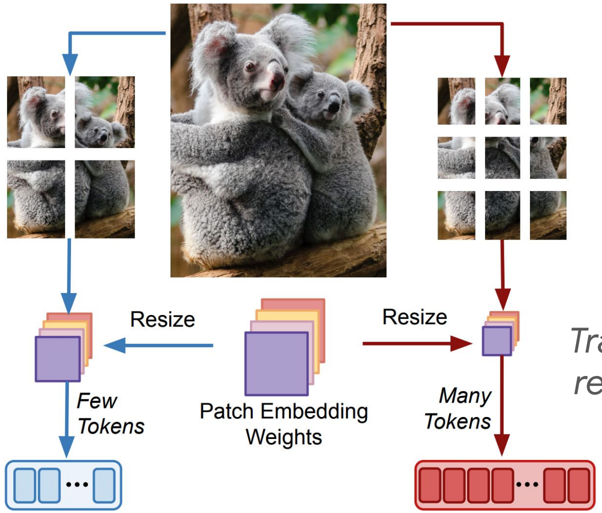
- **Model**: multimodal contrastive learning with SigLIP
- **Prior work**: only focuses on independent data selection, i.e. diagonals of the contrastive matrix
- **Intuition**: contrastive loss depends on entire matrix, and matrix is clearly non-diagonal!



Efficient scoring via online model approximation

- Data selection is expensive, cost scales linearly with amount of data rejected
- We use the FlexiVit architecture to score data at low resolution

Score data at low resolution

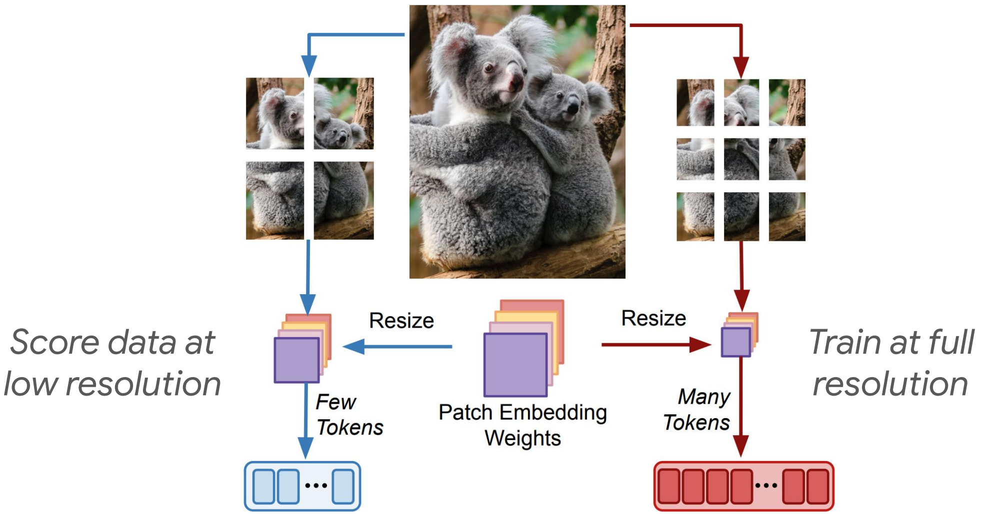


Train at full resolution

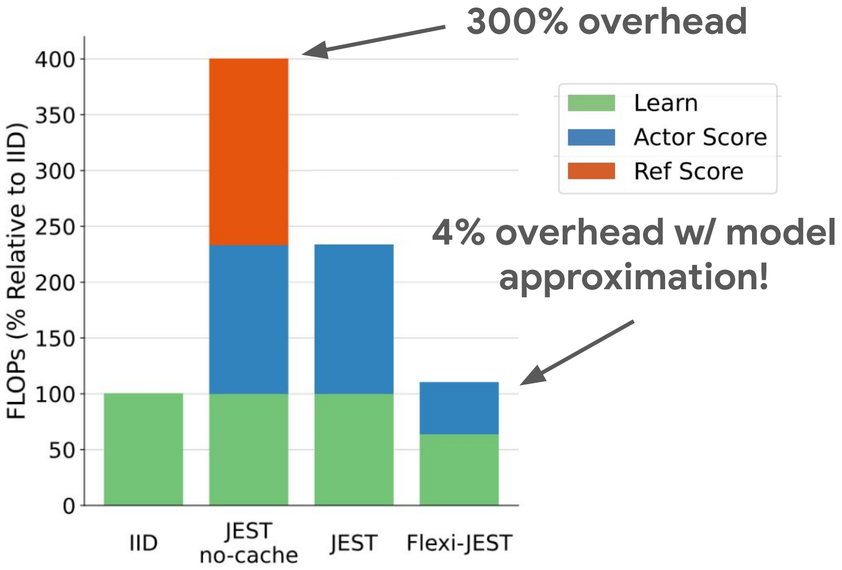
Beyer et al. (2023)

Efficient scoring via online model approximation

- Data selection is expensive, cost scales linearly with amount of data rejected
- We use the FlexiVit architecture to score data at low resolution

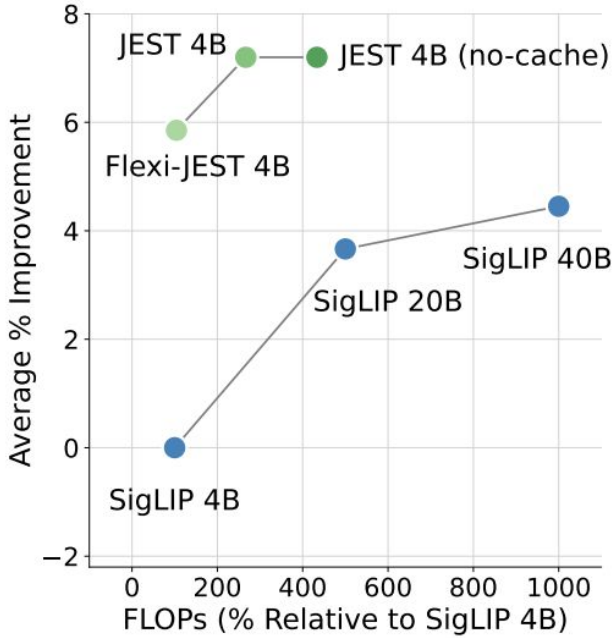


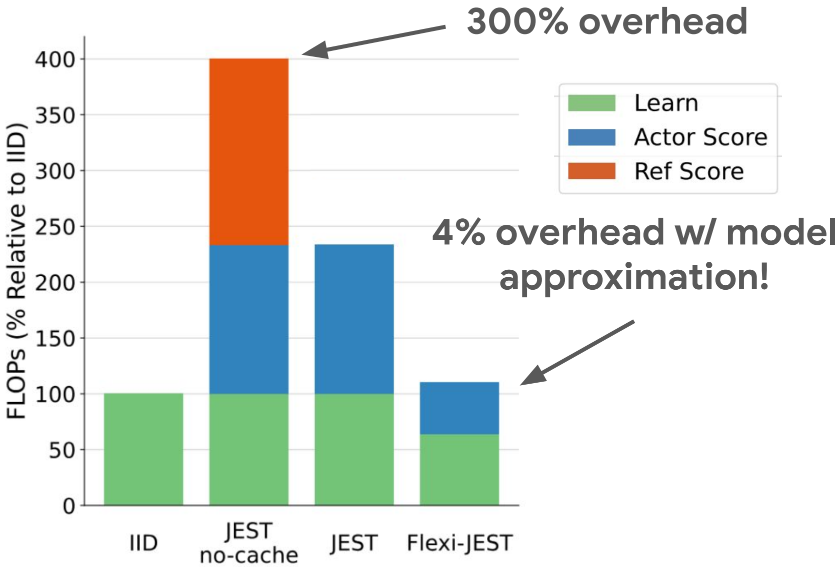
Beyer et al. (2023)



Efficient scoring via online model approximation

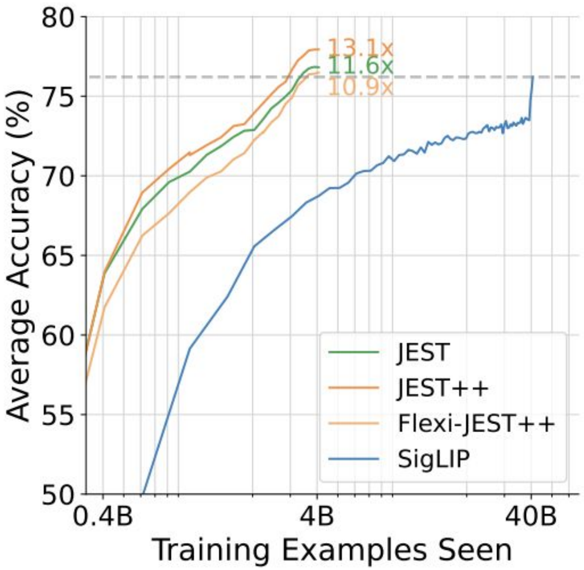
- Data selection is expensive, cost scales linearly with amount of data rejected
- We use the FlexiVit architecture to score data at low resolution





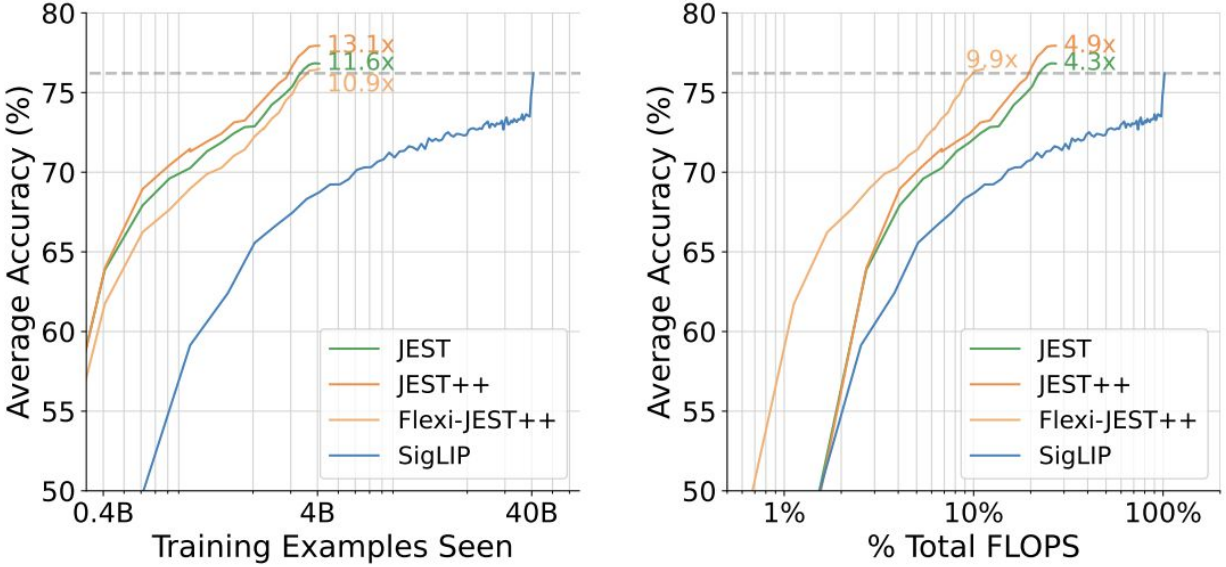
Joint Example Selection Accelerates Multimodal Learning

- Data selection is expensive, cost scales linearly with amount of data rejected
- We use the FlexiVit architecture to score data at low resolution



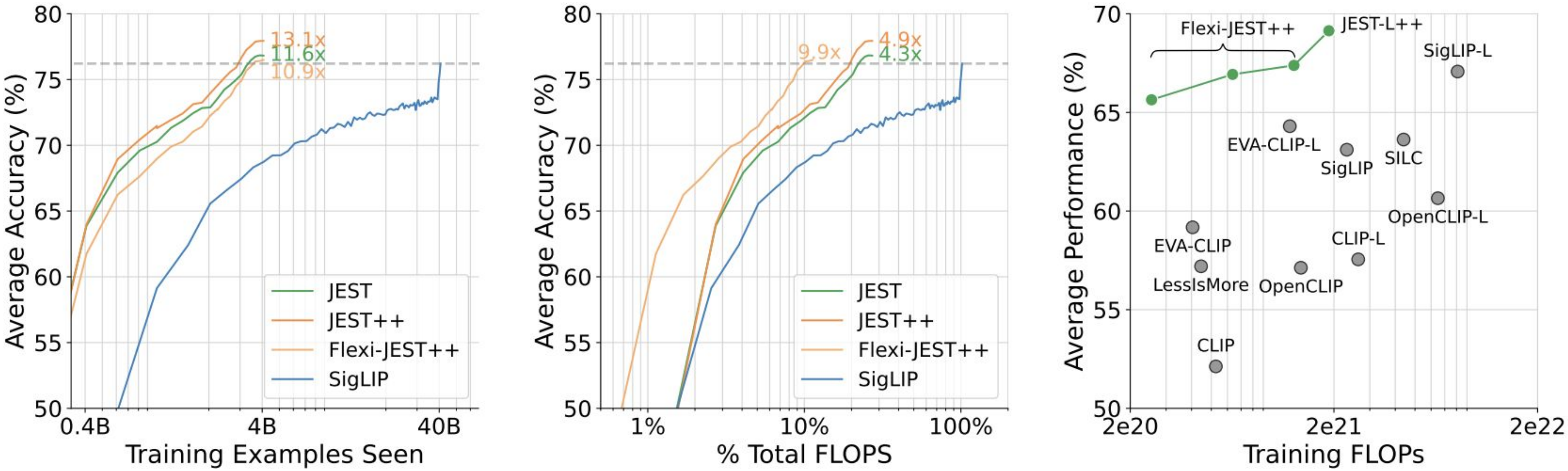
Joint Example Selection Accelerates Multimodal Learning

- Data selection is expensive, cost scales linearly with amount of data rejected
- We use the FlexiVit architecture to score data at low resolution



Joint Example Selection Accelerates Multimodal Learning

- Data selection is expensive, cost scales linearly with amount of data rejected
- We use the FlexiVit architecture to score data at low resolution



Model-based data curation meets self-supervised learning

Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding

⟶ builds a framework model-based data selection

- Which model-based criteria for data-selection?
- How to make data-selection tractable?

talfan@

nikparth@

rtanno@

Data Curation with Joint Example Selection Further Accelerates Multimodal Learning

⟶ applies this framework to multimodal contrastive SSL

- Contrastive SSL enables joint example selection (JEST)
- JEST radically accelerates multimodal learning (10x)

hamzamerzic@

schwarzjn@

shreyapa@

Model-based data curation meets self-supervised learning

Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding

⟶ builds a framework model-based data selection

- Which model-based criteria for data-selection?
- How to make data-selection tractable?

Data Curation with Joint Example Selection Further Accelerates Multimodal Learning

⟶ applies this framework to multimodal contrastive SSL

- Contrastive SSL enables joint example selection (JEST)
- JEST radically accelerates multimodal learning (10x)

ogle