universität freiburg

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

Talfan Evans*[1] Shreya Pathak*[1] Hamza Merzic*[1,2]
Jonathan Schwarz[1,3] Ryutaro Tanno[1] Olivier J. Henaff*[1]

*Equal technical contribution [1]Google DeepMind. [2]University College London. [3]Current affiliation: Harvard University, work done while at Google DeepMind.

Presentation by **Giacomo Mossio**

Supervised by **Simon Ging**

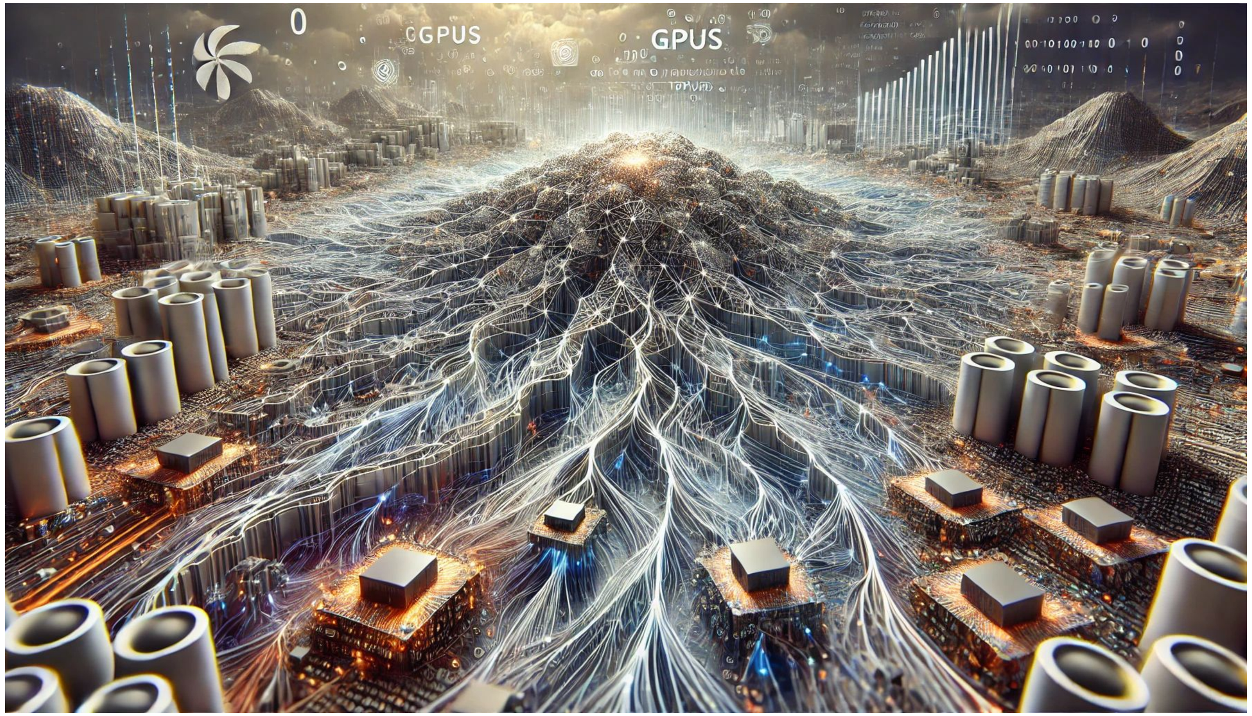
30.08.2024

Introduction

Large-Scale Models Training

Why Large-Scale Models:

- Revolutionizing Machine Learning.
- Applications in NLP, computer vision, etc.
- In general, the more data, the better.



[1]

Introduction

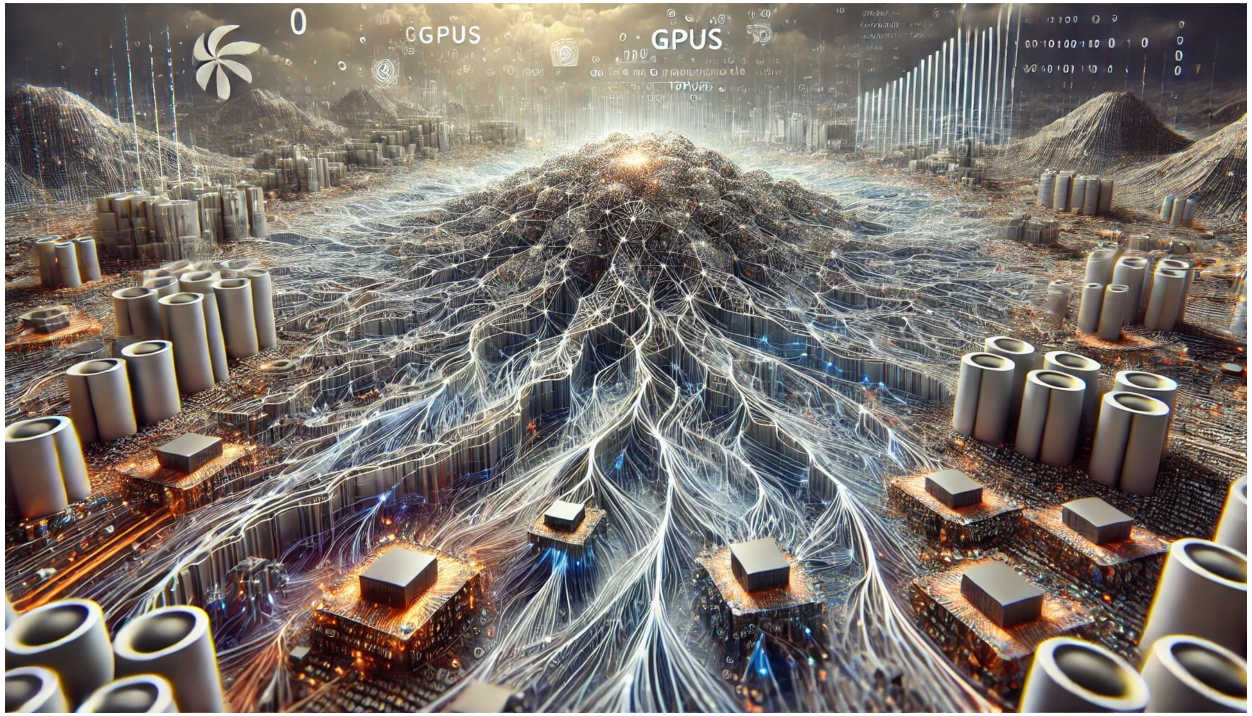
Large-Scale Models Training

Why Large-Scale Models:

- Revolutionizing Machine Learning.
- Applications in NLP, computer vision, etc.
- In general, the more data, the better.

Key Challenges:

- **Computational Requirements.**
- Scaling laws.
- Training with uniformly sampled data is slow.



[1]

Introduction

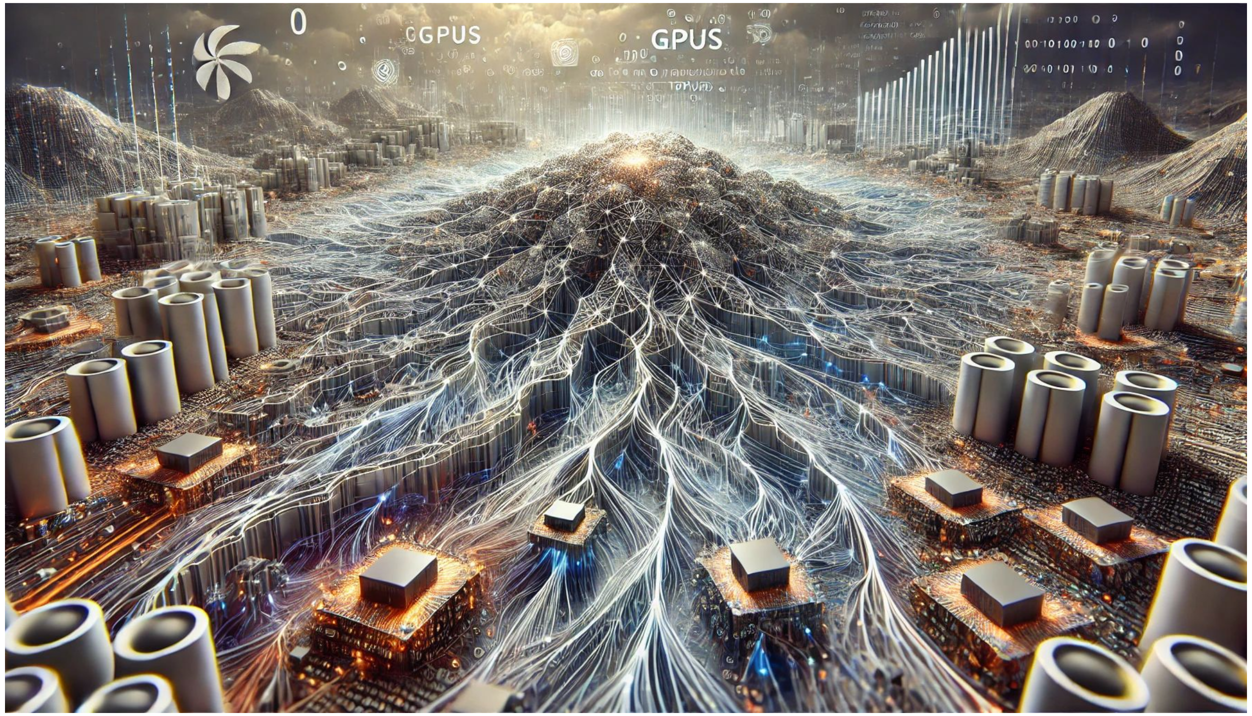
Large-Scale Models Training

Why Large-Scale Models:

- Revolutionizing Machine Learning.
- Applications in NLP, computer vision, etc.
- In general, the more data, the better.

Key Challenges:

- **Computational Requirements.**
- Scaling laws.
- Training with uniformly sampled data is slow.



[1]

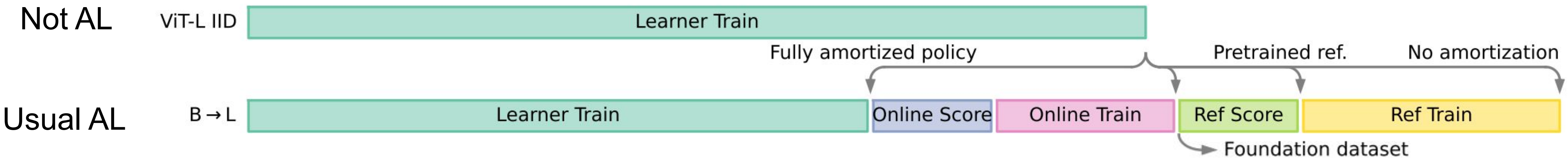
Can we reduce the computational costs of those type of models?

Objectives

Active Learning for Large-Scale Models pretraining

We want a data selection method that is:

1. General: robust to the choice of model and training task.
2. Scalable: works with large datasets and architectures.
3. **Compute-positive**: more compute efficient end-to-end than sampling training data randomly.



Agenda

1. Background

2. Related Work

3. Methods

4. Experiments

5. Results

6. Discussion

7. Conclusions and Future Work

Background

What is Active Learning (AL)

Challenges:

- Training models is expensive.
- Labelling data is expensive.
- Training is often redundant.

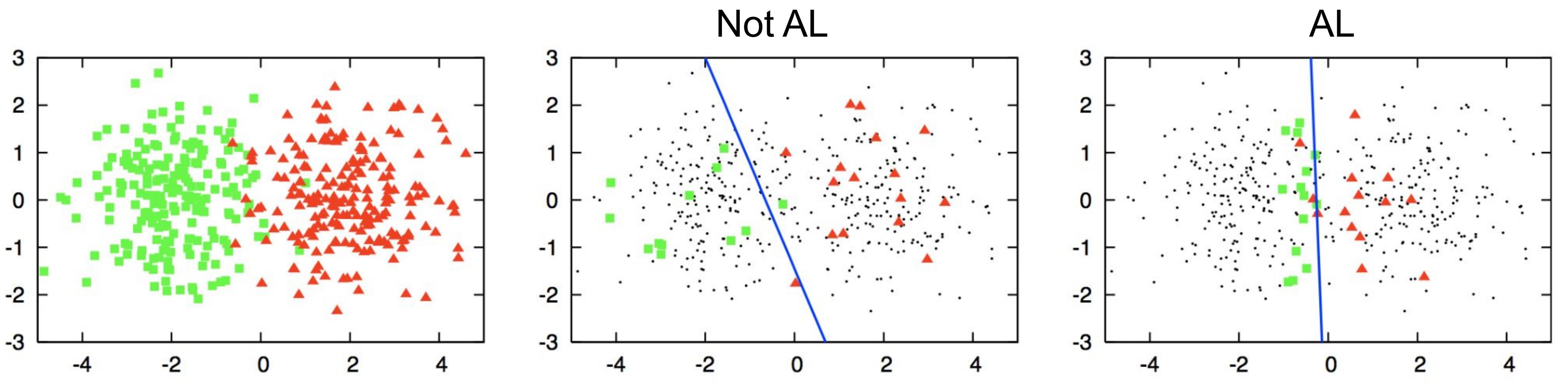
What is Active Learning (AL)

Challenges:

- Training models is expensive.
- Labelling data is expensive.
- Training is often redundant.

Solution:

- Some data points are more informative.
- AL estimates how valuable a data point is.
- AL aims at improving data efficiency.



[2]

Background

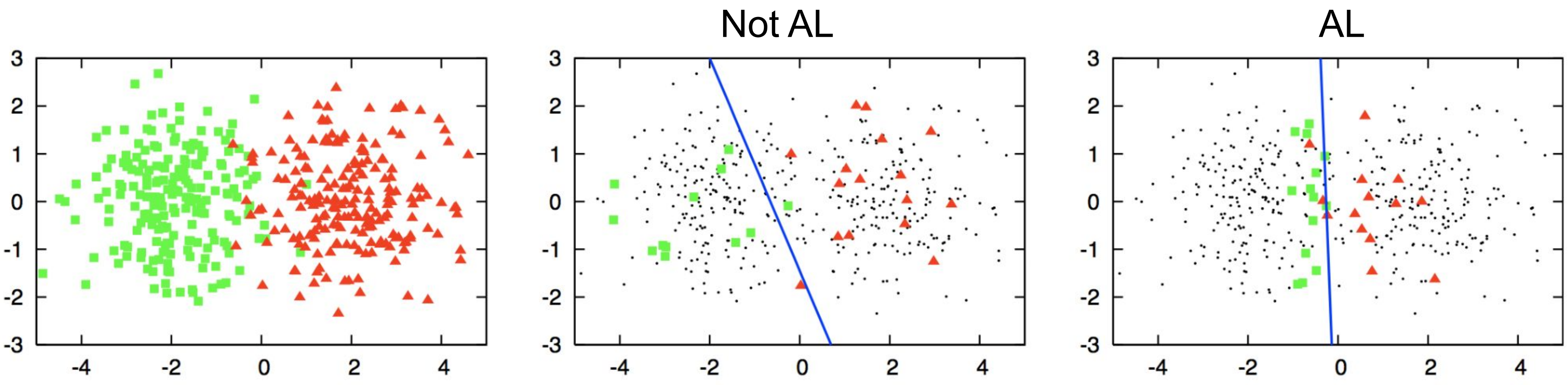
What is Active Learning (AL)

Challenges:

- Training models is expensive.
- Labelling data is expensive.
- Training is often redundant.

Solution:

- Some data points are more informative.
- AL estimates how valuable a data point is.
- AL aims at improving data efficiency.



[2]

How do we estimate the value of a data point?

Background

Active Learning

Various methods to calculate the **value** of a data point for our model [3]:

Background

Active Learning

Various methods to calculate the **value** of a data point for our model [3]:

1. **Uncertainty Sampling**: Model's predictions uncertainty.

Background

Active Learning

Various methods to calculate the **value** of a data point for our model [3]:

1. **Uncertainty Sampling**: Model's predictions uncertainty.
2. **Query-by-Committee**: Ensemble of models to find disagreement.

Active Learning

Various methods to calculate the **value** of a data point for our model [3]:

1. **Uncertainty Sampling**: Model's predictions uncertainty.
2. **Query-by-Committee**: Ensemble of models to find disagreement.
3. **Expected Model Change**: Greatest change in the model's parameters.

Background

Active Learning

Various methods to calculate the **value** of a data point for our model [3]:

1. **Uncertainty Sampling**: Model's predictions uncertainty.
2. **Query-by-Committee**: Ensemble of models to find disagreement.
3. **Expected Model Change**: Greatest change in the model's parameters.
4. **Expected Error Reduction**: Expected reduction of the overall error or loss of the model.

Background

Active Learning

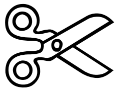
Various methods to calculate the **value** of a data point for our model [3]:

1. **Uncertainty Sampling**: Model's predictions uncertainty.
2. **Query-by-Committee**: Ensemble of models to find disagreement.
3. **Expected Model Change**: Greatest change in the model's parameters.
4. **Expected Error Reduction**: Expected reduction of the overall error or loss of the model.
5. **Core-Set Selection**: Subset of data points representing the diversity of the entire dataset.

Agenda

Related Work

Data Pruning



- Identifies and sub-selects data before training.
- Effective for small to medium datasets.
- Can be as expensive as learning in single-epoch training.

Related Work

Data Pruning

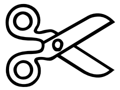
- Identifies and sub-selects data before training.
- Effective for small to medium datasets.
- Can be as expensive as learning in single-epoch training.

Online Active Learning

- Continuously filters data during training.
- Suitable for semi-infinite / single-epoch regime.
- Justifying efficiency gains vs. scoring costs.

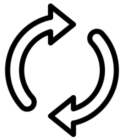
Related Work

Data Pruning



- Identifies and sub-selects data before training.
- Effective for small to medium datasets.
- Can be as expensive as learning in single-epoch training.

Online Active Learning



- Continuously filters data during training.
- Suitable for semi-infinite / single-epoch regime.
- Justifying efficiency gains vs. scoring costs.

Compute-efficient Data Selection

- Uses simple heuristics.
- Includes low-level image properties.
- May require domain specific knowledge or struggle with large-scale datasets.

Related Work

Reducible Holdout Loss (RHO-LOSS) Selection [4]

- **Online** Active Learning on web-scale data.
- Prioritizes points that are:
  - **Learnable** (**low noise**)
  - **Worth Learning** (**task-relevant**)
  - **Not Yet Learnt** (**non-redundant**)

Related Work

Reducible Holdout Loss (RHO-LOSS) Selection [4]

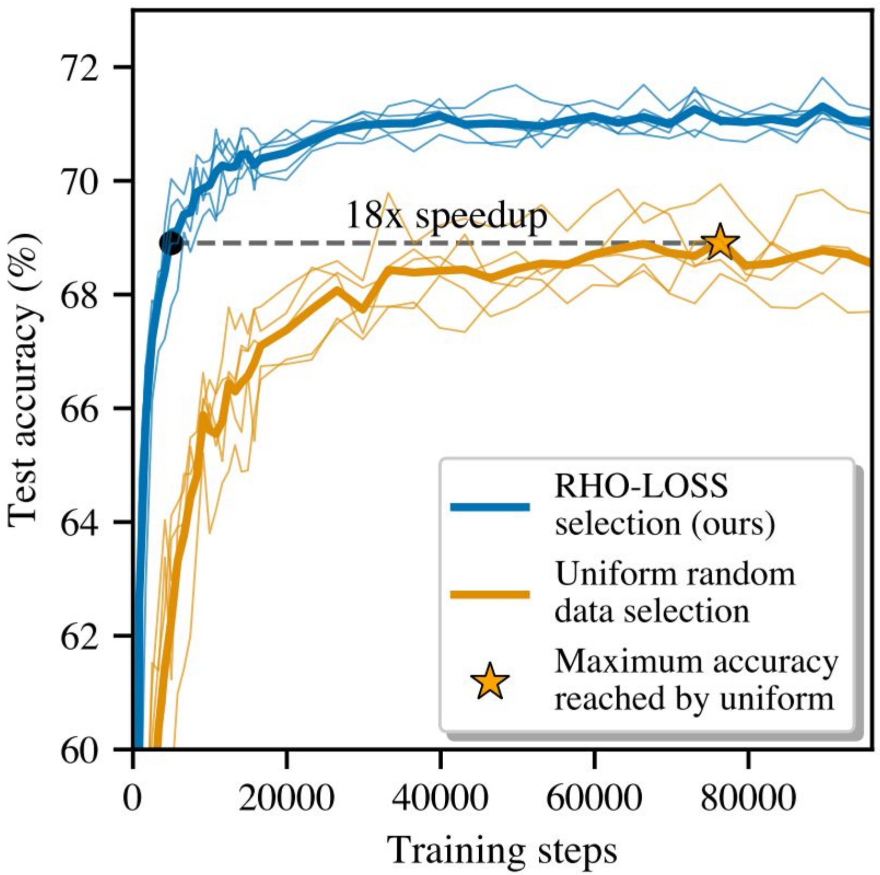
- **Online** Active Learning on web-scale data.
- Prioritizes points that are:
  - **Learnable** (**low noise**)
  - **Worth Learning** (**task-relevant**)
  - **Not Yet Learnt** (**non-redundant**)

- **Small reference model** trained on the ***holdout set***.

$$\underset{(x,y)\in B_t}{\arg\max} \overbrace{\underbrace{L[y \mid x; \mathcal{D}_{\mathrm{t}}]}_{\text{training loss}} - \underbrace{L[y \mid x; \mathcal{D}_{\mathrm{ho}}]}_{\text{irreducible holdout loss (IL)}}}^{\text{reducible holdout loss}}$$

Related Work

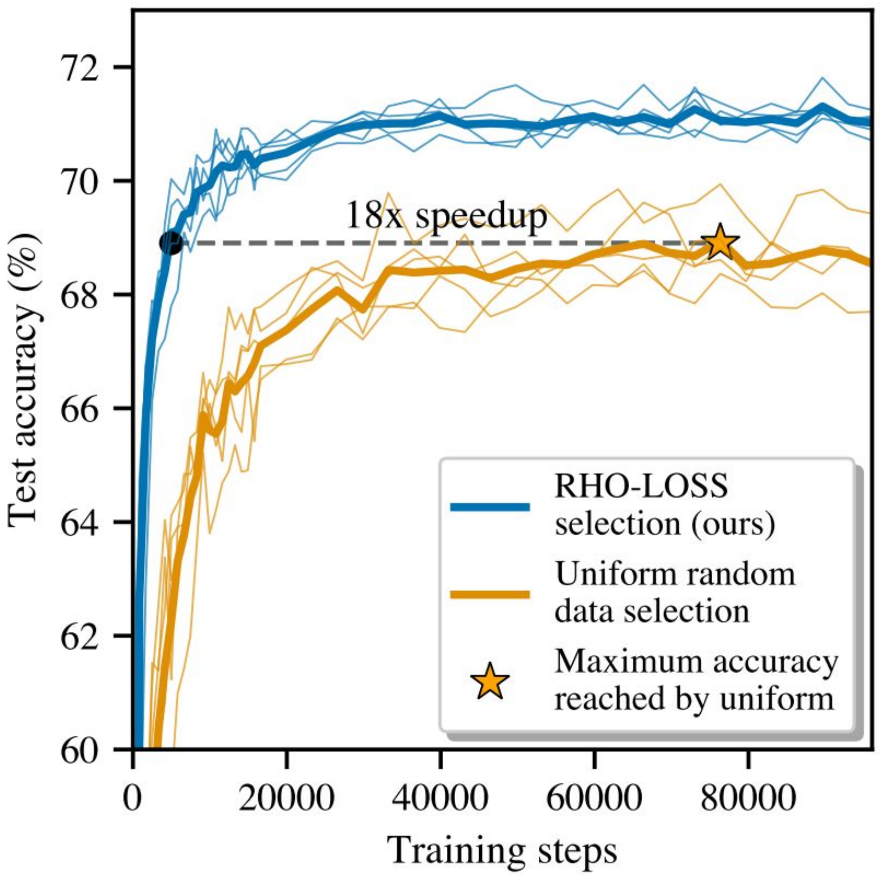
Reducible Holdout Loss (RHO-LOSS) Selection [4]



- Experiments on classification of web-scraped data (**Clothing-1M**).
- Impressive results on reducing **training steps.**

Related Work

Reducible Holdout Loss (RHO-LOSS) Selection [4]



- Experiments also on classification of web-scraped data (**Clothing-1M**).
- Impressive results on reducing **training steps.**

But we're not considering the **total compute speedup**

Agenda

Methods

Model-based prioritization

1. **Example difficulty:** Can be measured by training loss:
    - $s^{\text{hard}}(\boldsymbol{x}_i|\theta) = \ell(\boldsymbol{x}_i|\theta)$ for excluding trivial samples.
    - $s^{\text{easy}}(\boldsymbol{x}_i|\theta) = -\ell(\boldsymbol{x}_i|\theta)$ for excluding noisy samples.

Methods

Model-based prioritization

1. **Example difficulty:**
   - $s^{\text{hard}}(\boldsymbol{x}_i|\theta) = \ell(\boldsymbol{x}_i|\theta)$ for excluding trivial samples.
   - $s^{\text{easy}}(\boldsymbol{x}_i|\theta) = -\ell(\boldsymbol{x}_i|\theta)$ for excluding noisy samples.

2. **Example learnability:**
   - $$\begin{aligned} s^{\text{learn}}(\boldsymbol{x}_i|\theta^t, \theta^*) &= s^{\text{hard}}(\boldsymbol{x}_i|\theta^t) + s^{\text{easy}}(\boldsymbol{x}_i|\theta^*) \\ &= \ell(\boldsymbol{x}_i|\theta^t) - \ell(\boldsymbol{x}_i|\theta^*), \end{aligned}$$

   Where θ^t is the current learner and θ^* is a well trained model.

Methods

Compute-positive training

$$\underbrace{\Big(3F_{\text{learn}} + \rho F_{\text{act}}\Big)\beta + 3F_{\text{ref}}}_{\text{Active Learning}} < \underbrace{3F_{\text{learn}}}_{\text{IID}}$$

compute-positivity

F = cost of an inference pass.

We consider an inference pass as ⅓ of gradient update.

F_{learn} = cost of learner model

F_{act} = cost of scoring

F_{ref} = cost of reference model

β = ratio of samples used compared to IID.

ρ = number of examples scored per training example.

Methods

Compute-positive training

$$\underbrace{\Big(3F_{\text{learn}} + \rho F_{\text{act}}\Big)\beta + 3F_{\text{ref}}}_{\text{Active Learning}} < \underbrace{3F_{\text{learn}}}_{\text{IID}}$$

compute-positivity

- **RHO** learnability scoring: $F_{act} = F_{ref} + \boldsymbol{F_{learn}}$ → inference through a reference model **and the learner**.

Methods

Compute-positive training

$$\underbrace{\Big(3F_{\text{learn}} + \rho F_{\text{act}}\Big)\beta + 3F_{\text{ref}}}_{\text{Active Learning}} < \underbrace{3F_{\text{learn}}}_{\text{IID}}$$

compute-positivity

- **RHO** learnability scoring: $F_{act} = F_{ref} + \boldsymbol{F_{learn}}$ → inference through a reference model **and the learner**.

- **This method:** We replace F_{learn} for scoring with a proxy model ($\boldsymbol{F_{online}}$) with same size of F_{ref} →

$$F_{act} = F_{ref} + \boldsymbol{F_{online}} = \boldsymbol{2F_{ref}}$$

Methods

Algorithm

θ_l is the big main **learner** model.
θ_o is the small **online** model used to find hard samples.
θ_r is the small pretrained **reference** model used to exclude easy samples.
B is the batch size.
b is the sub-batch (usually ½ B).

Methods

Algorithm

θ_l is the big main **learner** model.
θ_o is the small **online** model used to find hard samples.
θ_r is the small pretrained **reference** model used to exclude easy samples.
B is the batch size.
b is the sub-batch (usually ½ B).

while training **do**
  $X \sim \mathcal{D}$, where $|X| = B$ ▷ Sample IID
  $S = \ell_{\text{act}}(X|\theta_o) - \ell_{\text{act}}(X|\theta_r)$ ▷ Get scores
  $I \sim \text{SoftMax}(S)$, where $|I| = b$ ▷ Sample indices
  $Y = X[I]$ ▷ Collect sub-batch
  $\theta_l \leftarrow \text{Adam}[\nabla_{\theta_l} \ell_{\text{learn}}(Y|\theta_l)]$ ▷ Update learner model
  $\theta_o \leftarrow \text{Adam}[\nabla_{\theta_o} \ell_{\text{learn}}(Y|\theta_o)]$ ▷ Update online model
end while

Methods

Losses

1) ClassAct **(Visual classification)**: Standard cross entropy for scoring and learner

$$\ell_{\mathrm{CE}}(x_i|\theta) = -\sum_{c=1}^{C} y_{ic} \log p_{ic}(x_i;\theta)$$

Methods

Losses

1) ClassAct **(Visual classification)**: Standard cross entropy for scoring and learner

$$\ell_{\text{CE}}(x_i|\theta) = -\sum_{c=1}^{C} y_{ic} \log p_{ic}(x_i;\theta)$$

2) ActiveCLIP (**Multimodal learning):** Contrastive loss for learner and dot-product similarity img-txt for scoring

$\boldsymbol{z}_i^{\text{im}}$ = image embeddings
$\boldsymbol{z}_i^{\text{txt}}$ = text embeddings

$$\ell_{\text{learn}}^{\text{im,txt}}(\boldsymbol{x}_i|\theta) = -\log \frac{\exp(\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_i^{\text{txt}})}{\sum_j \exp(\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_j^{\text{txt}})}$$

$$\ell_{\text{learn}} = \ell_{\text{learn}}^{\text{im,txt}} + \ell_{\text{learn}}^{\text{txt,im}}$$

$$\ell_{\text{act}}(\boldsymbol{x}_i|\theta) = -\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_i^{\text{txt}}$$

Methods

Losses

1) ClassAct **(Visual classification)**: Standard cross entropy for scoring and learner

$$\ell_{\text{CE}}(x_i|\theta) = -\sum_{c=1}^{C} y_{ic} \log p_{ic}(x_i;\theta)$$

2) ActiveCLIP (**Multimodal learning):** Contrastive loss for learner and dot-product similarity img-txt for scoring

$\boldsymbol{z}_i^{\text{im}}$ = image embeddings
$\boldsymbol{z}_i^{\text{txt}}$ = text embeddings

$$\ell_{\text{learn}}^{\text{im,txt}}(\boldsymbol{x}_i|\theta) = -\log \frac{\exp(\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_i^{\text{txt}})}{\sum_j \exp(\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_j^{\text{txt}})}$$

$$\ell_{\text{learn}} = \ell_{\text{learn}}^{\text{im,txt}} + \ell_{\text{learn}}^{\text{txt,im}}$$

$$\ell_{\text{act}}(\boldsymbol{x}_i|\theta) = -\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_i^{\text{txt}}$$

3) ActiveSigLIP (**Multimodal learning**): Sigmoid loss for learner and dot-product similarity img-txt for scoring

$$\ell_{\text{learn}}(x_i|\theta) = -\sum_{c=1}^{C} [y_{ic} \log \sigma(z_{ic}) + (1 - y_{ic}) \log(1 - \sigma(z_{ic}))]$$

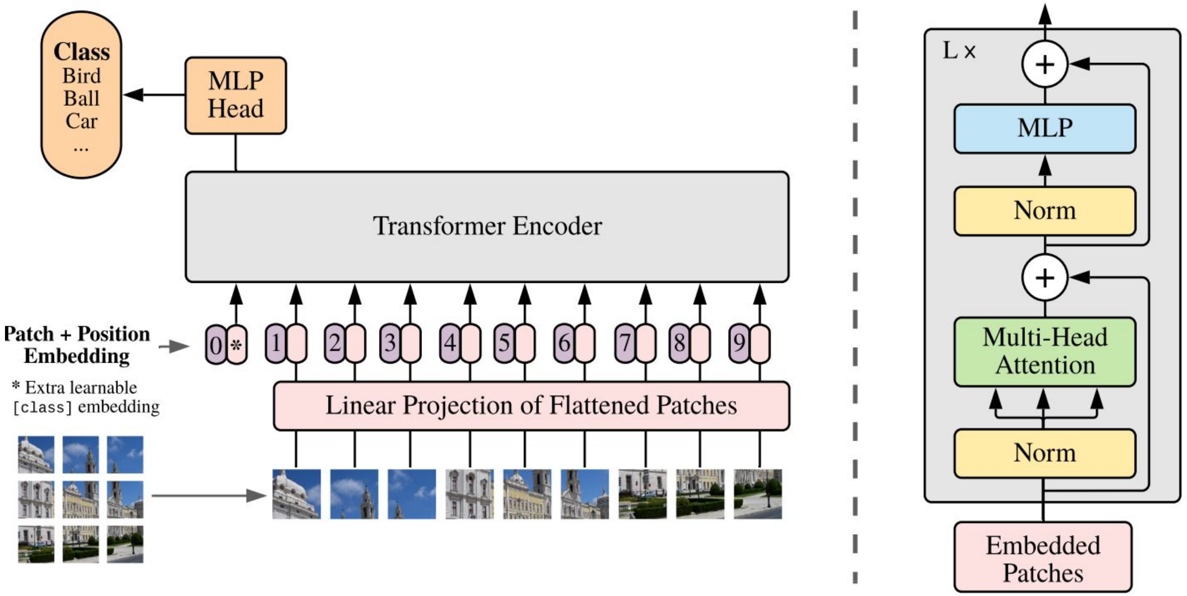
$$\ell_{\text{act}}(\boldsymbol{x}_i|\theta) = -\boldsymbol{z}_i^{\text{im}} \cdot \boldsymbol{z}_i^{\text{txt}}$$

Agenda

Experiments

Visual Classification (ClassAct)

Vision Transformer (ViT) [5]

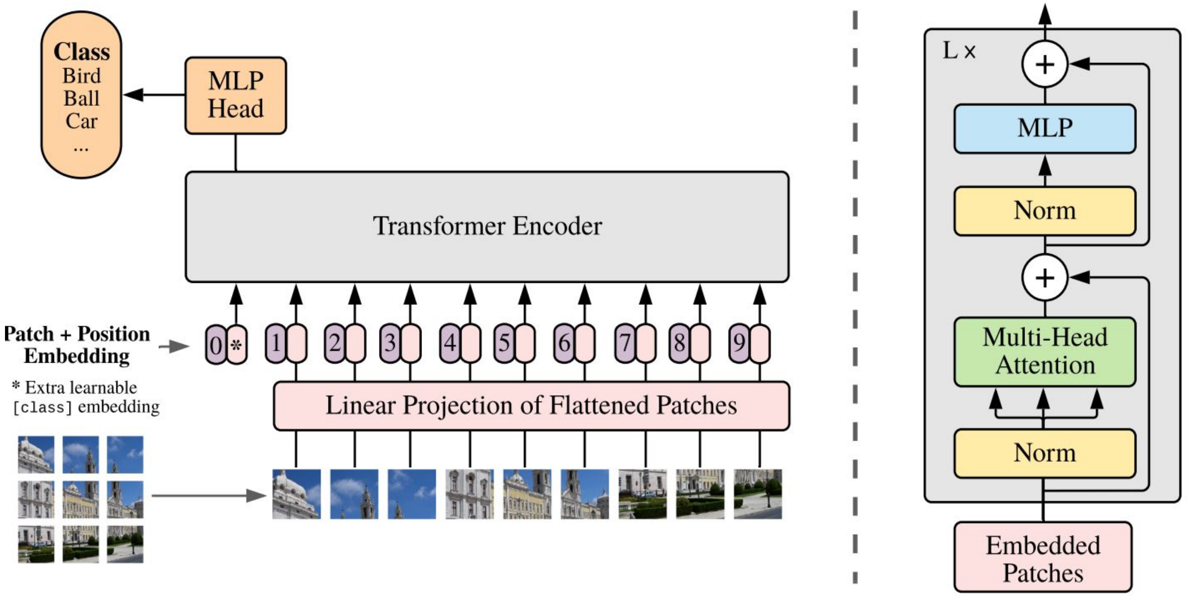


- Image patches as sequence tokens.
- Self-attention for global context.
- Scalable for large datasets.

Experiments

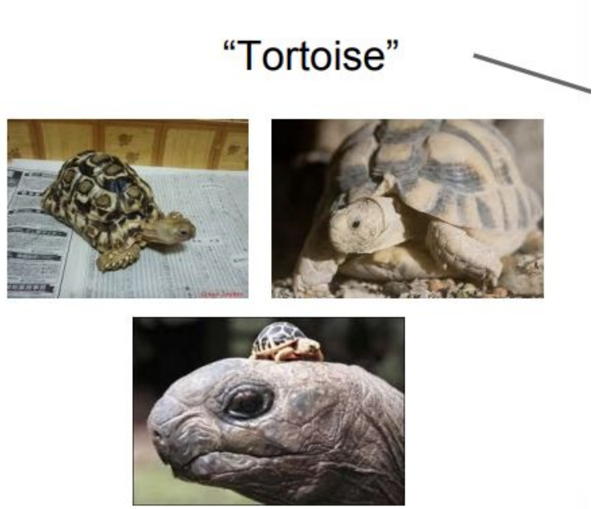
Visual Classification (ClassAct)

Vision Transformer (ViT) [5]



- Image patches as sequence tokens.
- Self-attention for global context.
- Scalable for large datasets.

JTF 300M [6]

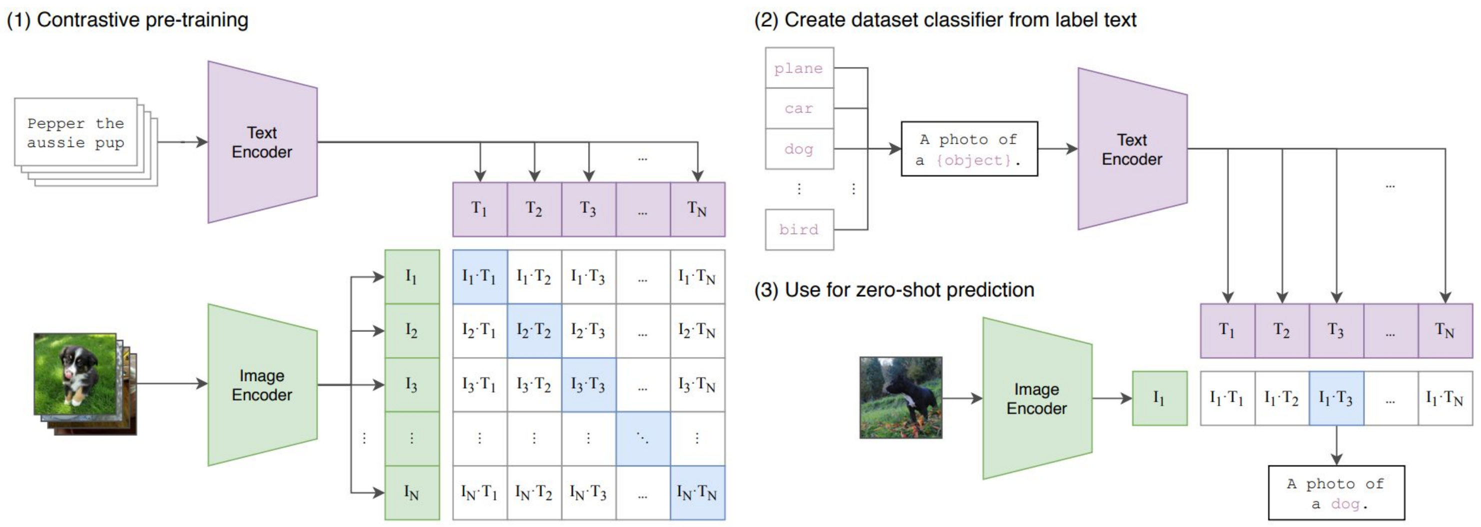


- Internal Google dataset.
- Large-scale image classification.
- 1B labels for the 300M images.

Experiments

Multimodal Learning (ActiveCLIP)

CLIP [7]

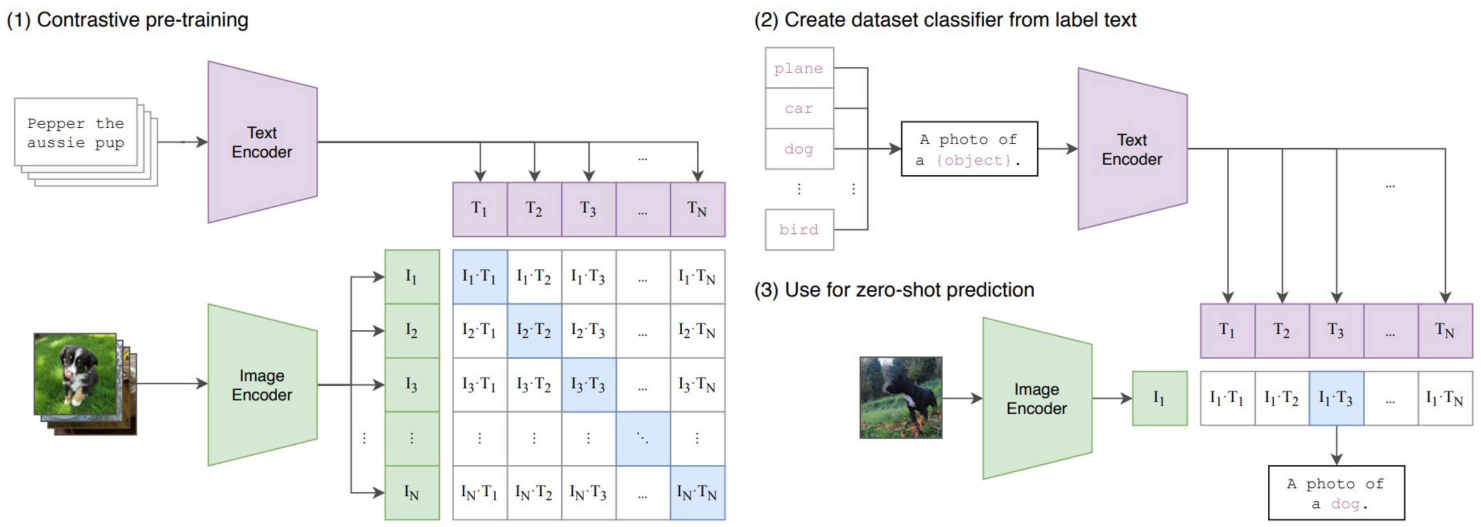


- **CLIP** links images and text embeddings (contrastive learning).

Experiments

Multimodal Learning (ActiveCLIP)

CLIP [7]

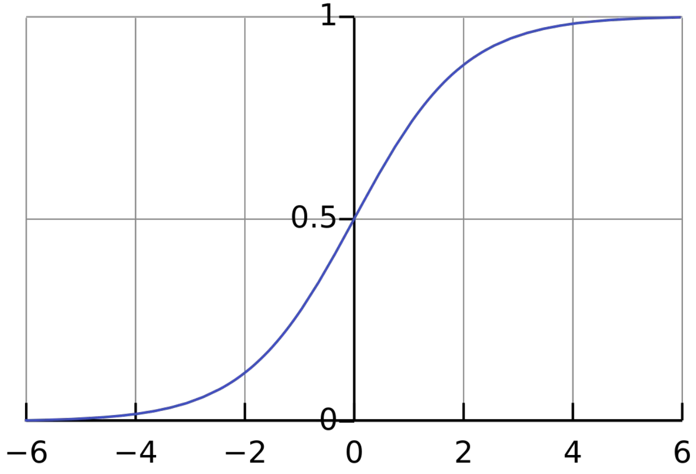


- **CLIP** links images and text embeddings (contrastive learning).
- **ALIGN**: Multimodal dataset of image-text pairs.
- **LTIP**: Curated dataset with diverse, clean data.
- **JTF 300M**: Can be used for contrastive learning too.

Experiments

Multimodal Learning (ActiveSigLIP)

SigLIP [8]

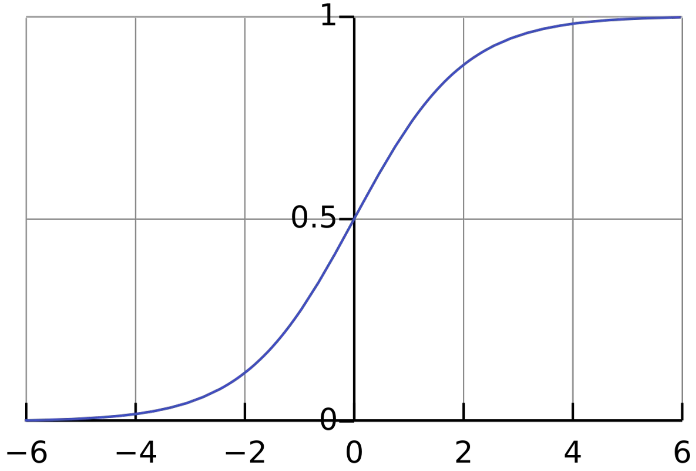


- ***ActiveSigLIP*** based on Sigmoid loss for Language-Image Pre-training **(SigLIP)**.
- Learns matching probabilities between image-text pairs.

Experiments

Multimodal Learning (ActiveSigLIP)

SigLIP [8]



WebLI [9]

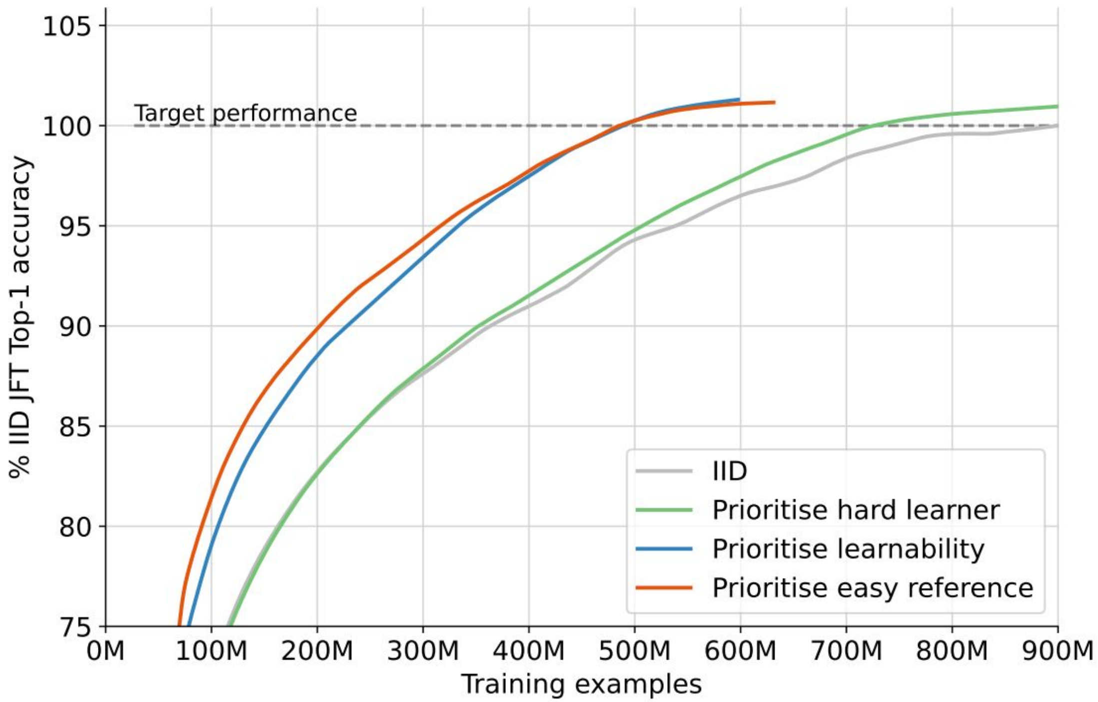
- ***ActiveSigLIP*** based on Sigmoid loss for Language-Image Pre-training **(SigLIP)**.
- Learns matching probabilities between image-text pairs.

- Large-scale **multilingual** image-language dataset from the **web**.
- **109 languages** and **10 billions** of image-text and image-OCR pairs.
- high-quality subset of 1 billion examples.

Agenda

Results

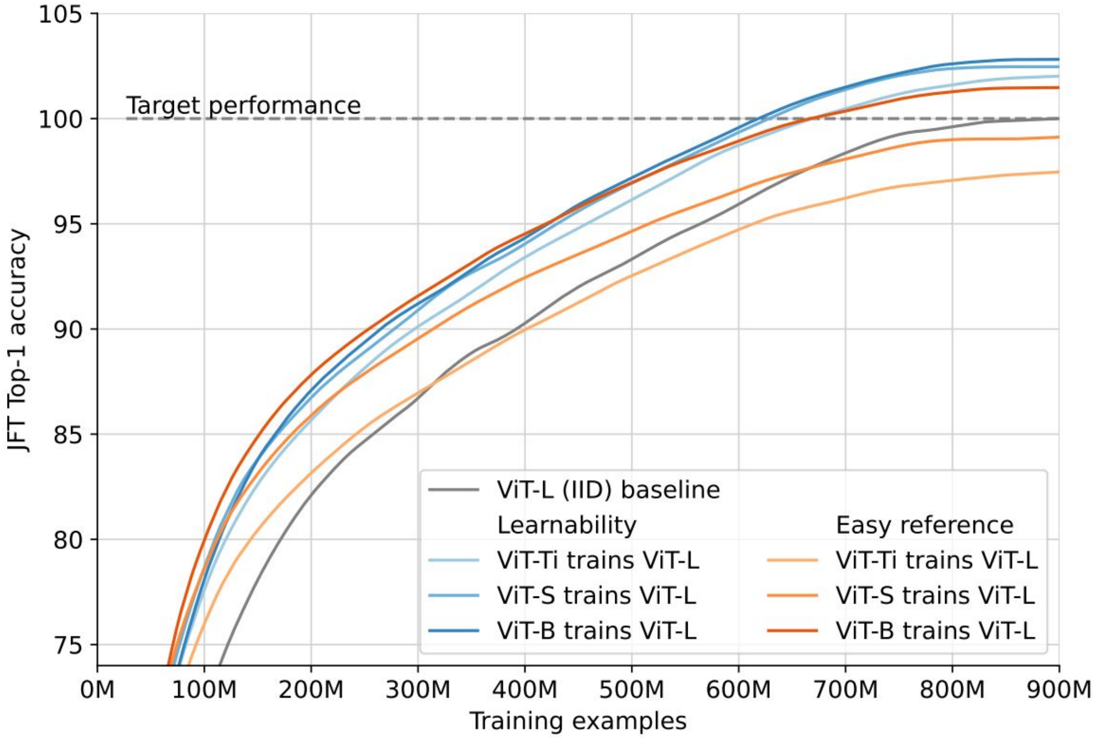
Data-selection criteria



- Hard learner is not a good criteria.
- Strongly **compute-negative** because learner additional inference.

Results

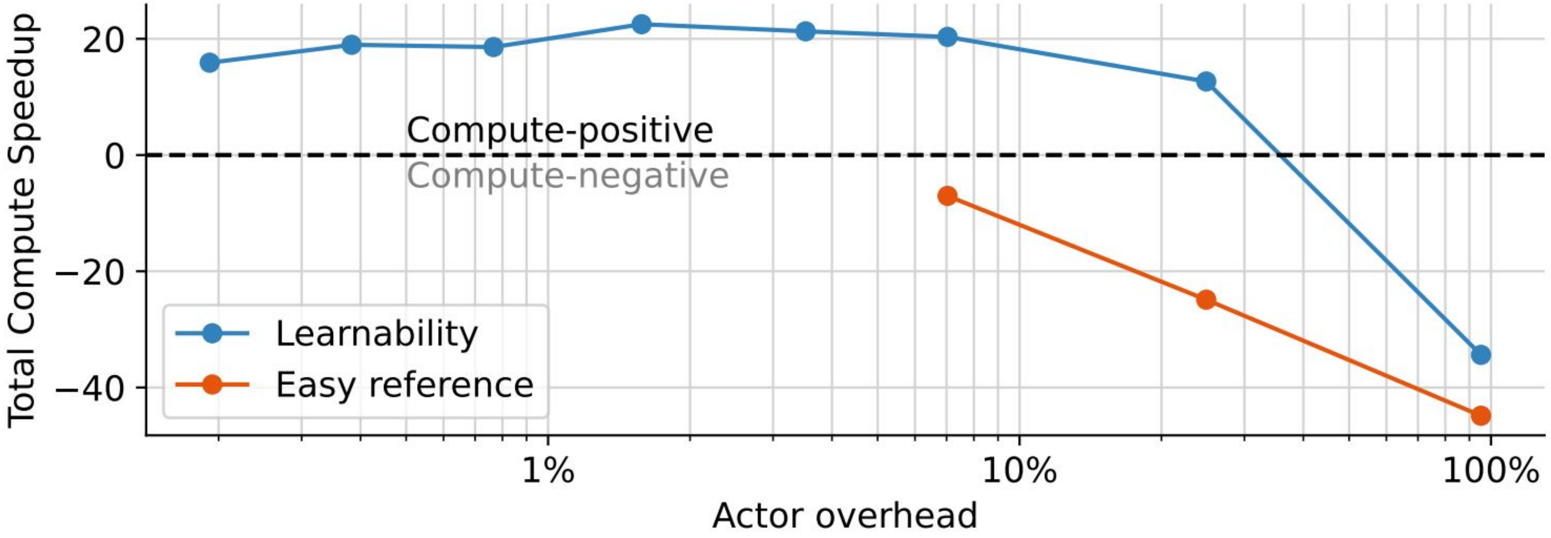
Performances across actors scales



- Easy reference model sizes affect a lot performances.
- Learnability **small** models are good.

Results

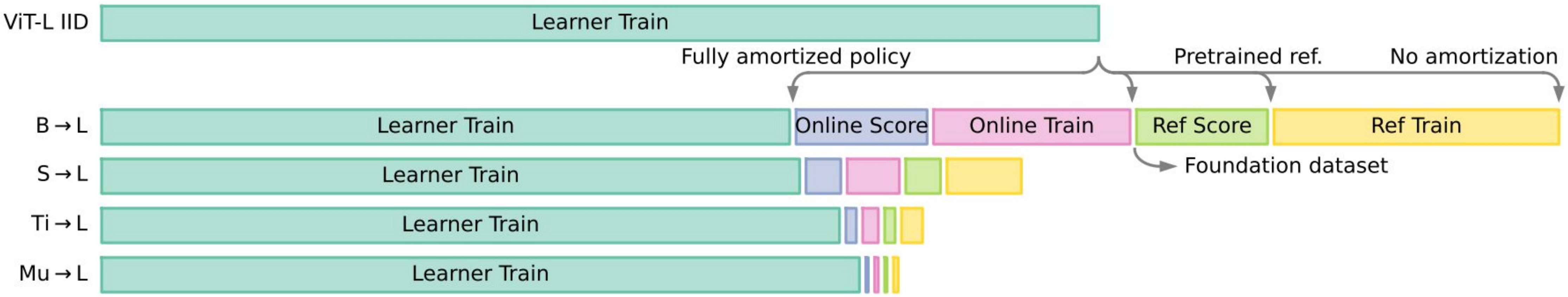
Total compute speedup across actors scales



- Actor overhead = additional FLOPs to compute the scores.
- **Compute-positivity** by scaling down actors models.

Results

Amortizing the cost of data selection

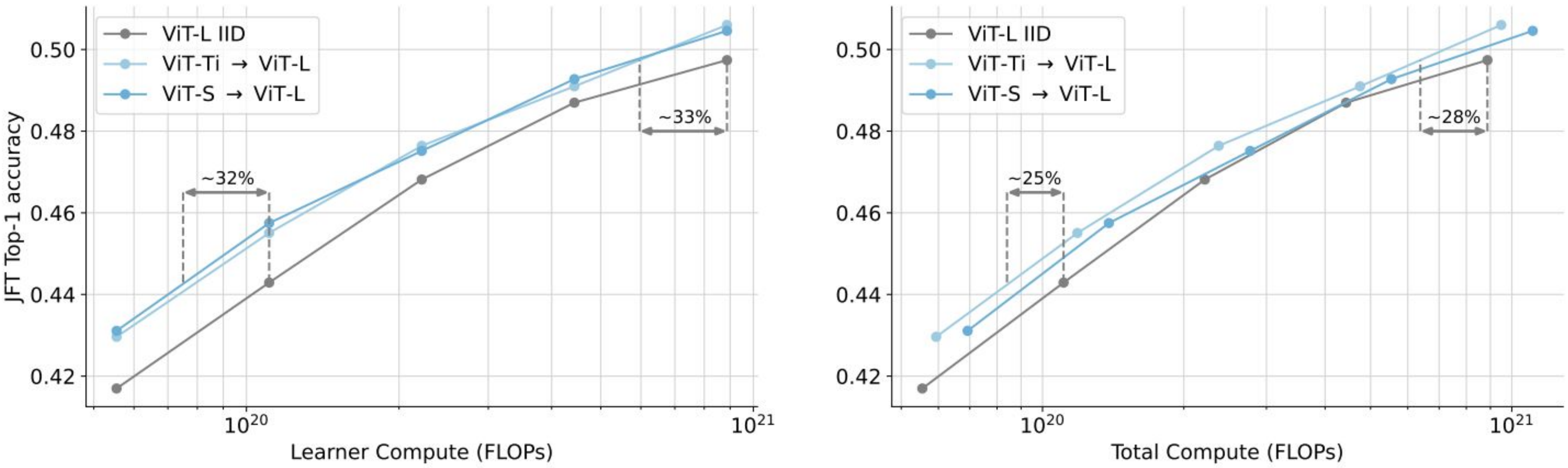


Possible amortizations:

- **Off-the-shelf** reference model (yellow cost).
- Scores assigned once to a ‘**foundation dataset’** (lime cost).

Results

Scaling laws for active learning



- First paper showing a **general model-based** method that shift scaling laws in our favour.

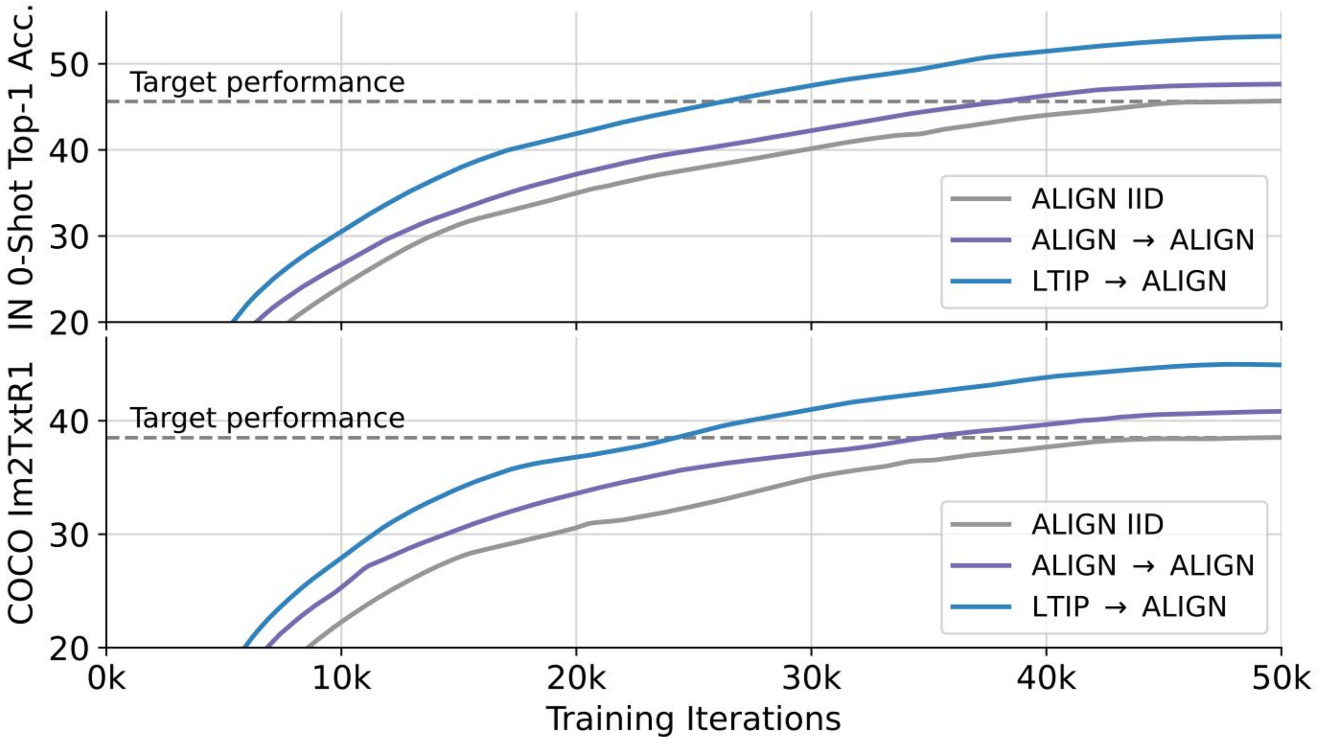
Results

Reference model types

Method	ViT model capacity			Reference Type	Speed-up	
	Reference	Online	Learner		Learner speedup	Compute speedup
ViT-B IID			B		0%	0%
RHO	Tiny	B	B	Held-out, fixed	0%	- 79%
ClassAct-HO	Tiny	Tiny	B	Held-out, fixed	18%	3%
ClassAct	Tiny	Tiny	B	In-domain, fixed	18%	3%
ClassAct-Online	Tiny	Tiny	B	Trained online	17%	2%

Results

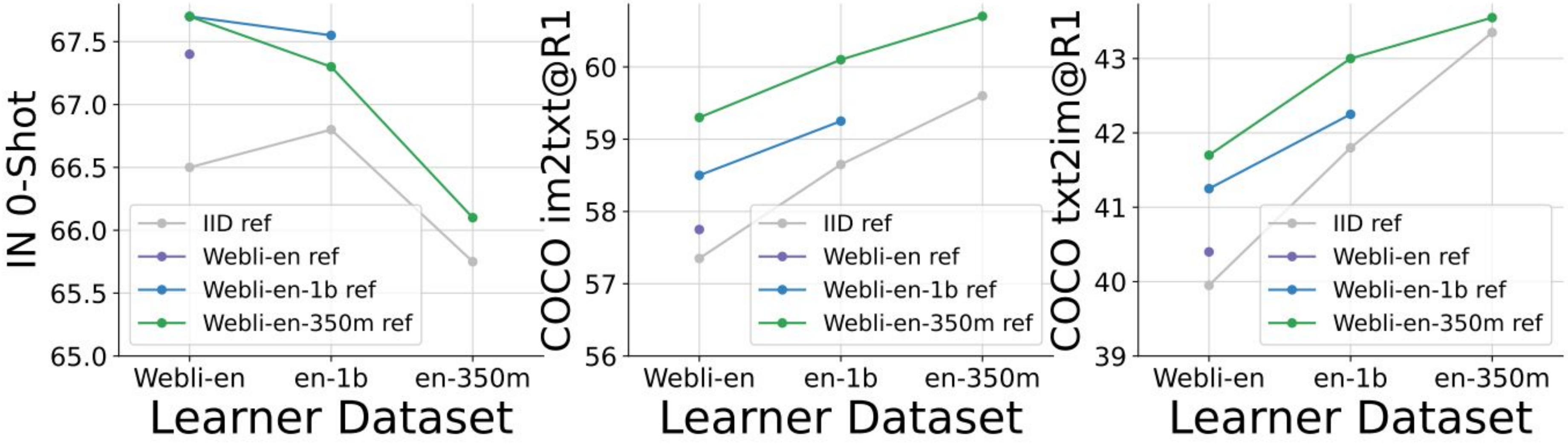
ActiveCLIP - Multimodal training



- Reference models trained on **related but distinct** datasets are better.

Results

ActiveSigLIP - Multimodal training



- **Small curated** datasets are better references.

Results

Multimodal training

		IN-1K	COCO	
Method	Train ex.	ZS Top-1	im2txt	txt2im
CLIP	13B	68.3	52.4	33.1
EVA-CLIP	3B+2B	69.7†		
ActiveCLIP	3B	**71.3**	**57.7**	**43.0**
OpenCLIP	34B	70.2	59.4	42.3
EVA-CLIP	8B+2B	74.7†	58.7	42.2
ActiveCLIP	8B	**72.2**	**60.7**	**44.9**
SigLIP	3B	72.1	60.7	42.7
ActiveSigLIP	3B	72.0	**63.5**	**45.3**

- ***ActiveCLIP*** outperforms models trained with the same or more data

Agenda

Discussion

Strengths:

1. **Compute efficiency**: Data efficiency reduces total costs.

2. **Scalability**: well-suited for large datasets and complex models.

3. **Generalization**: The approach works with different models and tasks.

Discussion

Strengths:

1. **Compute efficiency**: Data efficiency reduces total costs.
2. **Scalability**: well-suited for large datasets and complex models.
3. **Generalization**: The approach works with different models and tasks.

Limitations:

1. **Compute Costs**: The compute-positivity margin is not that large.
2. **Reference Models**: The approach depend on the selection of the reference model.
3. **Infrastructure**: Online learning with large infrastructure adds complexity.

Agenda

Conclusions and Future Work

Conclusions:

1. **Data efficiency:** Active Learning is good for large scale applications too.

2. **Method**: Learnability scores with two small proxy models.

3. **Results**: First method to be compute-positive and not model-dependent.

Conclusions and Future Work

Conclusions:

1. **Data efficiency:** Active Learning is good for large scale applications too.
2. **Method**: Learnability scores with two small proxy models.
3. **Results**: First method to be compute-positive and not model-dependent.

Future Works:

1. **Filtering ratio**: Always experimented filtering only 50% of the data.
2. **New domains**: Extend it to language, video and generative models.

References

[1] Image created using OpenAI's ChatGPT with DALL-E

[2] image from https://www.datacamp.com/tutorial/active-learning

[3] Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

[4] Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Kornblith, S., Fort, S., ... & Farhadi, A. (2022). Prioritized training on points that are learnable, worth learning, and not yet learnt. arXiv. https://arxiv.org/abs/2206.07137

[5] Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2020). Sigmoid Loss for Language-Image Pre-Training. arXiv. https://arxiv.org/abs/2010.11929

[6] image from https://paperswithcode.com/dataset/jft-300m

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

[8] Zhai, X., Steiner, A., & Beyer, L. (2023). Sigmoid loss for language-image pre-training. NeurIPS.

[9] Zhai, X., Lin, Z., Liu, H., & Zhang, K. (2022). PaLI: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2210.09793.

Thank you!

Questions

?