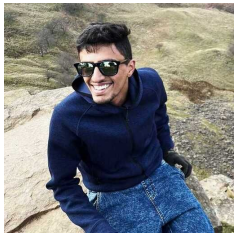
EBERHARD KARLS UNIVERSITÄT TÜBINGEN • ellis European Laboratory for Learning and Intelligent Systems • UNIVERSITY OF CAMBRIDGE • UNIVERSITY OF OXFORD

No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance

EBERHARD KARLS UNIVERSITÄT TÜBINGEN • ellis European Laboratory for Learning and Intelligent Systems • UNIVERSITY OF CAMBRIDGE • UNIVERSITY OF OXFORD

Vishaal Udandarao

Ameya Prabhu



Adhiraj Ghosh

Yash Sharma

Philip Torr

Adel Bibi

Samuel Albanie

Matthias Bethge

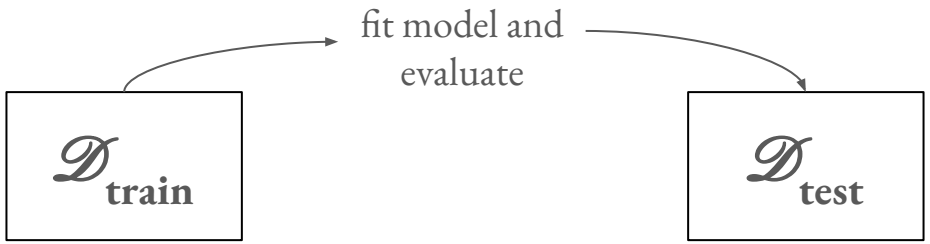
What is generalization?

What is generalization?

In the olden days (classic ML) ...

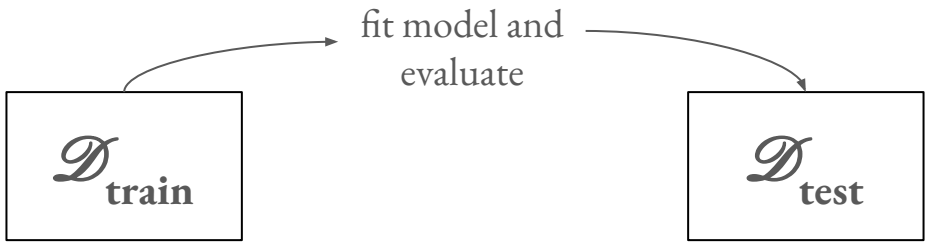
What is generalization?

In the olden days (classic ML) ...



What is generalization?

In the olden days (classic ML) ...



Small-scale and controllable!
We know what's in here.

Usually held-out!
Totally unseen samples.

What is generalization?

Then, training sets started getting **bigger**...

What is generalization?

Then, training sets started getting **bigger**...

Annotating them started getting **costlier**...

What is generalization?

Then, training sets started getting **bigger**...

Annotating them started getting **costlier**...

Solution: **“Zero-shot Learning/Generalization”**

What is “zero-shot” generalization?

Importance of Semantic Representation: Dataless Classification

Ming-Wei Chang, Lev Ratinov, Dan Roth and **Vivek Srikumar**
Department of Computer Science
University of Illinois at Urbana-Champaign
{mchang21, ratinov2, danr, vsrikum2}@uiuc.edu

AAAI’08

Zero-data Learning of New Tasks

Hugo Larochelle and **Dumitru Erhan** and **Yoshua Bengio**
Université de Montréal
Montréal, Québec
{larocheh,erhandum,bengioy}@iro.umontreal.ca

AAAI’08

Zero-Shot Learning with Semantic Output Codes

Mark Palatucci
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
mpalatuc@cs.cmu.edu

Dean Pomerleau
Intel Labs
Pittsburgh, PA 15213
dean.a.pomerleau@intel.com

Geoffrey Hinton
Computer Science Department
University of Toronto
Toronto, Ontario M5S 3G4, Canada
hinton@cs.toronto.edu

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
tom.mitchell@cs.cmu.edu

NeurIPS’09

What is “zero-shot” generalization?

And perhaps the most famous one…

What is “zero-shot” generalization?

And perhaps the most famous one…

Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer

Christoph H. Lampert Hannes Nickisch Stefan Harmeling
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
{firstname.lastname}@tuebingen.mpg.de

What is “zero-shot” generalization?

And perhaps the most famous one…

Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer

Christoph H. Lampert Hannes Nickisch Stefan Harmeling
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
{firstname.lastname}@tuebingen.mpg.de

Basic premise: Training and testing classes are disjoint!

What is “zero-shot” generalization?

And perhaps

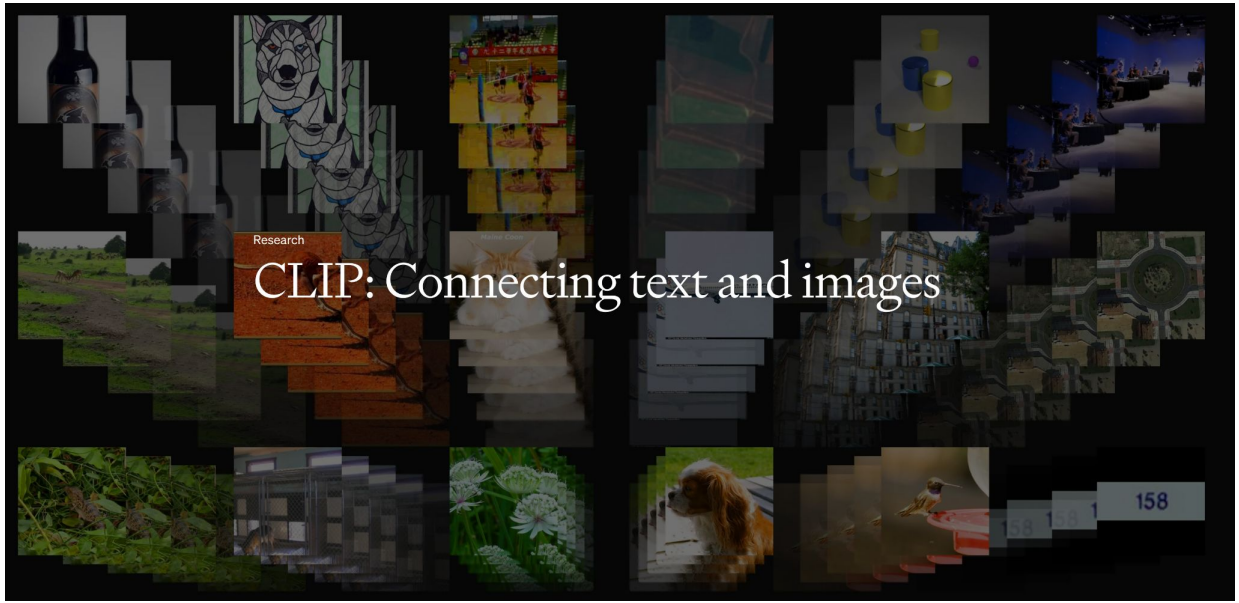
Hilde Kuehne
@HildeKuehne

In the old 'attribute-based' zero-shot times, the assumption was that you have seen furry animals, brown animals, and animals living in the wood, but technically you have not seen any bear (if that's e.g. your test class) ... the problem is that now you don't know if a bear is really in your dataset (resp. if you check, you will find a lot of bears mentioned on webscale data) ... this is what the no-zero-shot paper is about (huggingface.co/papers/2404.04...) and what we also found in our VL-Taboo paper when we looked at attributes (arxiv.org/abs/2209.06103)

So I would argue the best we can claim is that VL models were not trained on the downstream datasets (but might have the class knowledge), but even this can be questioned now if we follow arxiv.org/abs/2404.04125

Along comes CLIP!

Cut to 2021...



Along comes CLIP!

Cut to 2021...

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

Along comes CLIP!

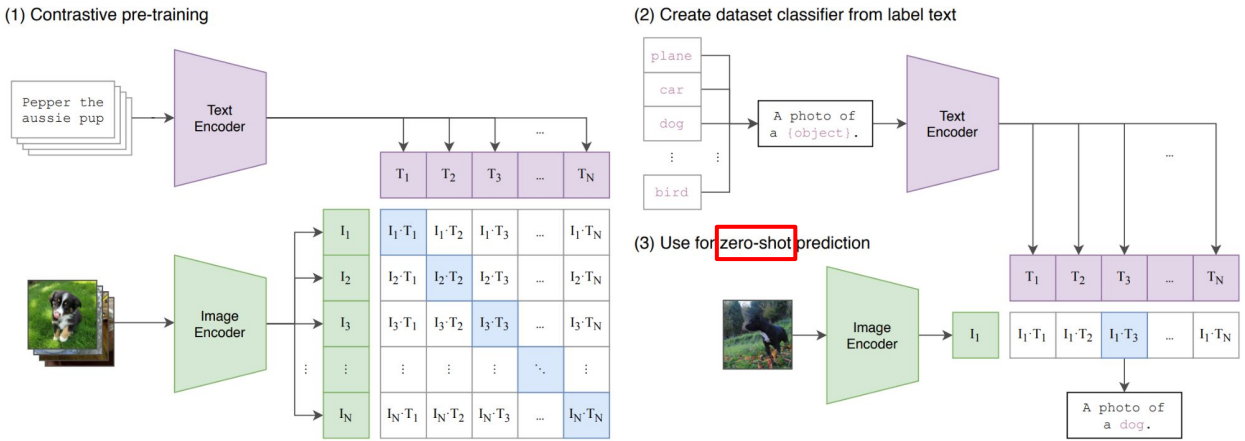


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Along comes CLIP!

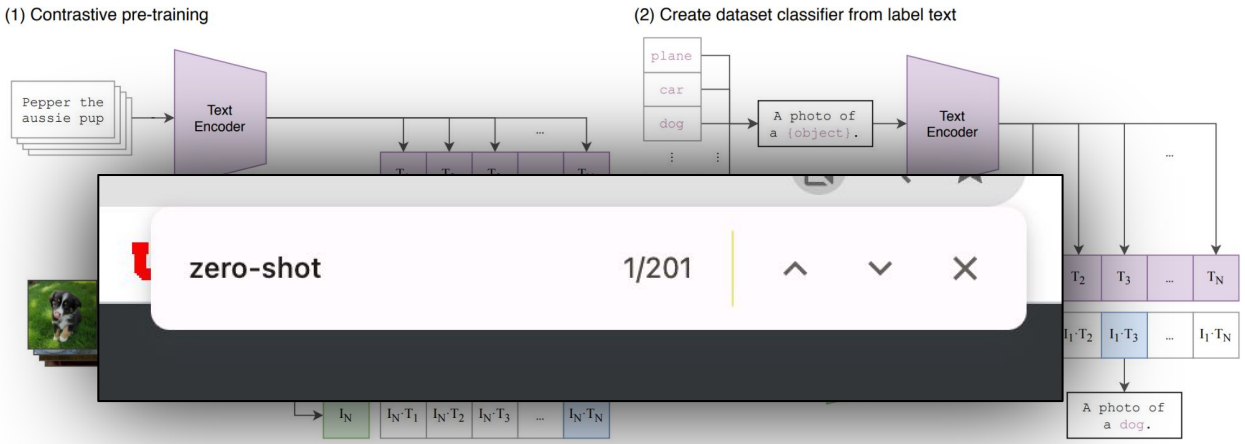
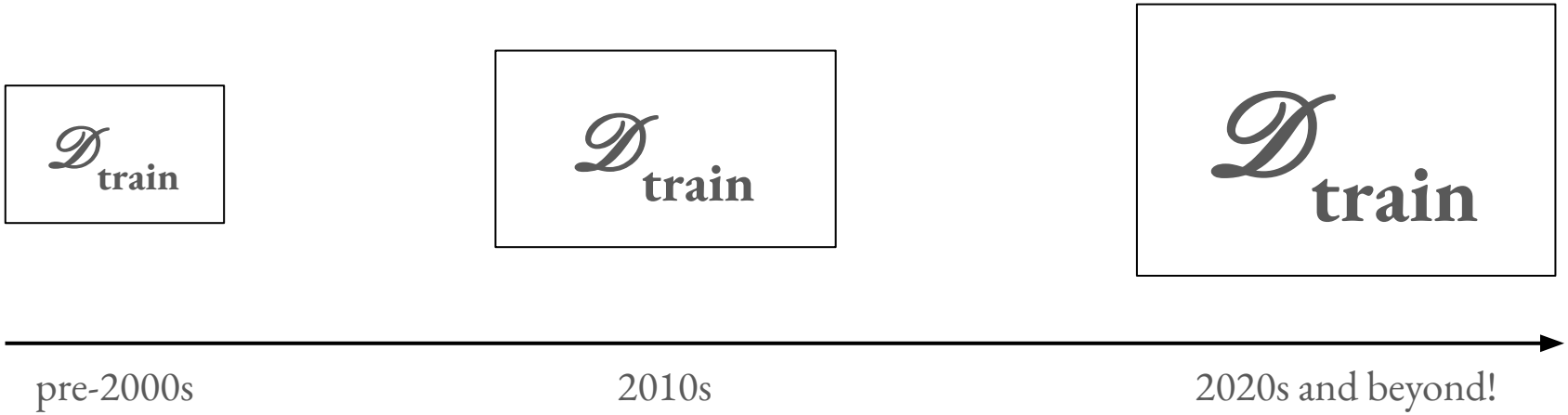
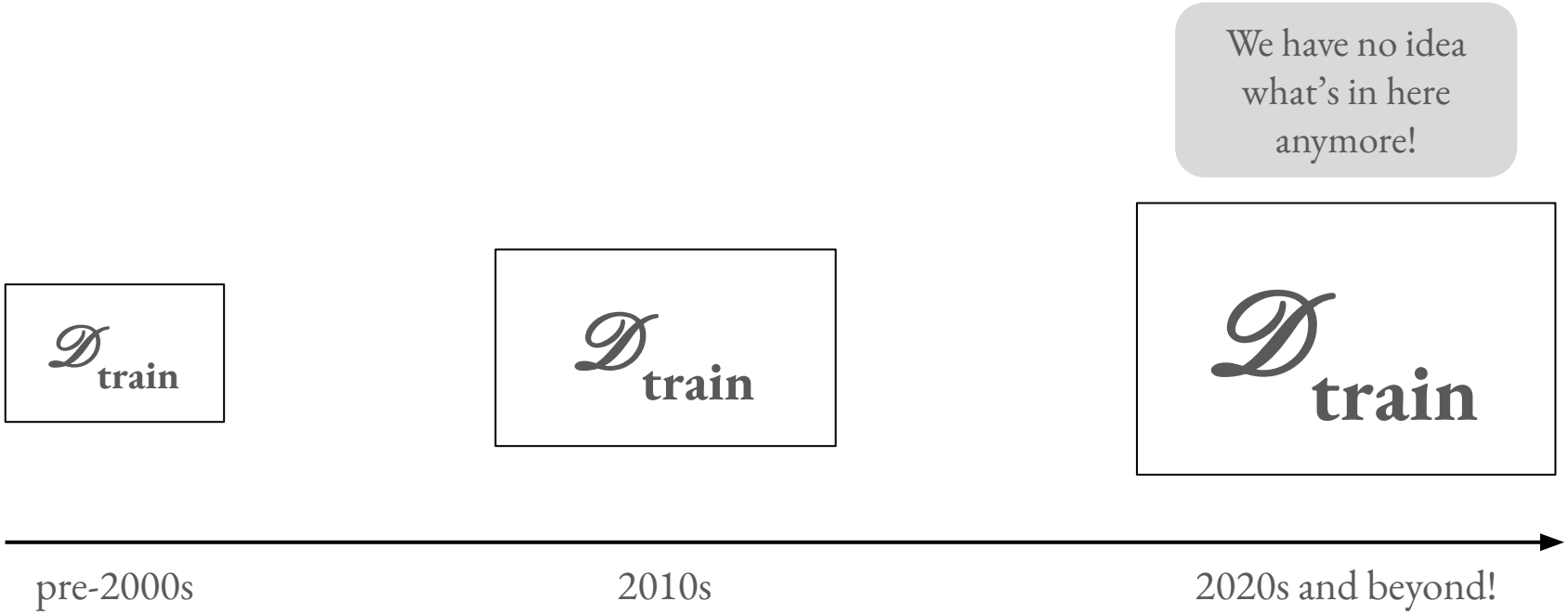


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

The term “zero-shot” has diffused over time – Opaque and massive training datasets!



The term “zero-shot” has diffused over time – Opaque and massive training datasets!



Our Goal

Take a sober look at the **“zero-shot” generalization** of multimodal models.

A simple methodology to evaluate “zero-shot” generalization

Understand what’s here

A simple methodology to evaluate “zero-shot” generalization

Understand what’s here

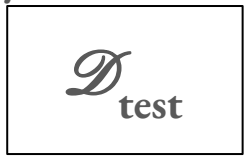
and

A simple methodology to evaluate “zero-shot” generalization

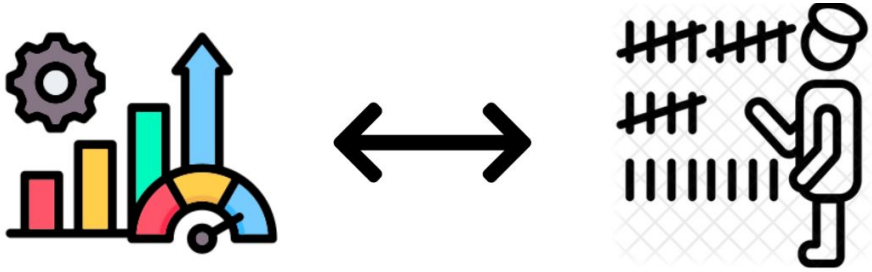
Understand what’s here

and

Verify with what’s here



Our key research question

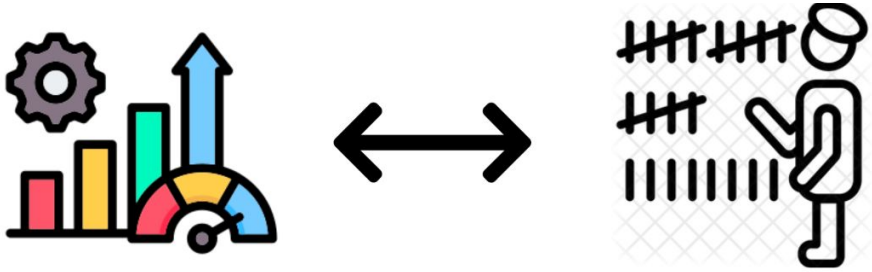


Model Performance

Concept Frequency

How is the performance of multimodal models on downstream concepts influenced by the frequency of these concepts in their pretraining datasets?

Our key research question



Model Performance

Concept Frequency

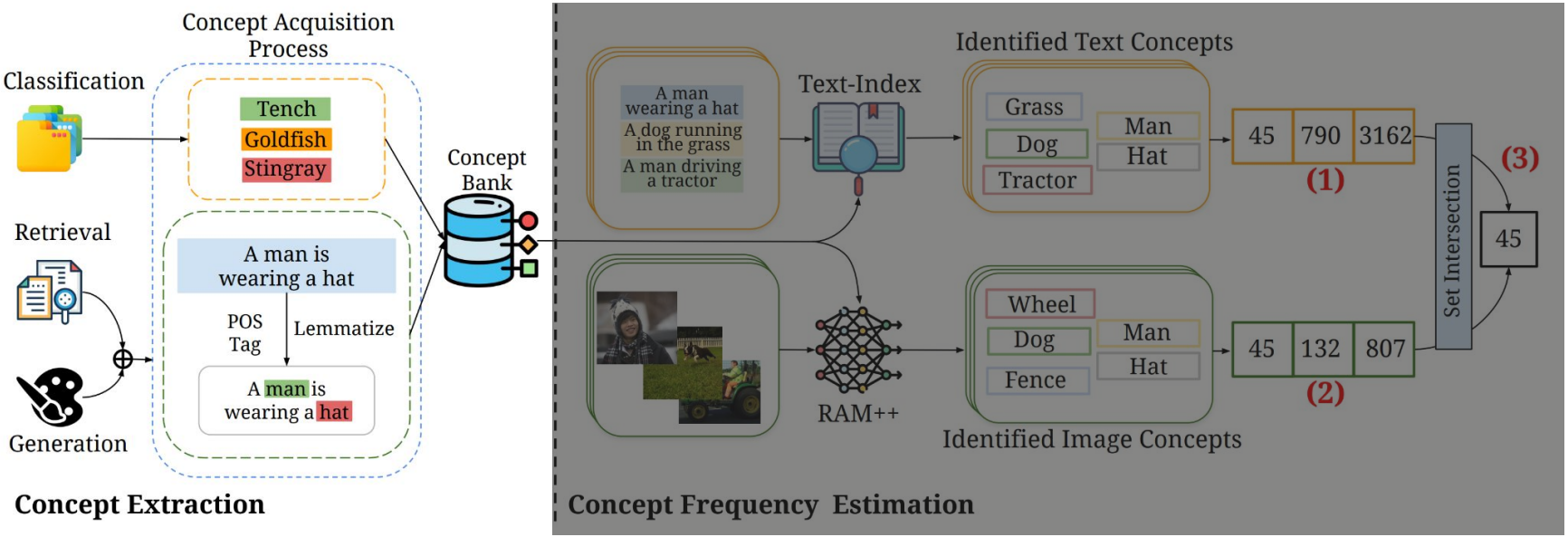
Pretraining Datasets:
CC-3M
CC-12M
YFCC-15M
LAION-400M
LAION-Aesthetics

Downstream Datasets:
17 “zero-shot” classification
2 image-text retrieval
8 text-to-image generation

Tested Models:
10 CLIP models
24 text-to-image gen models

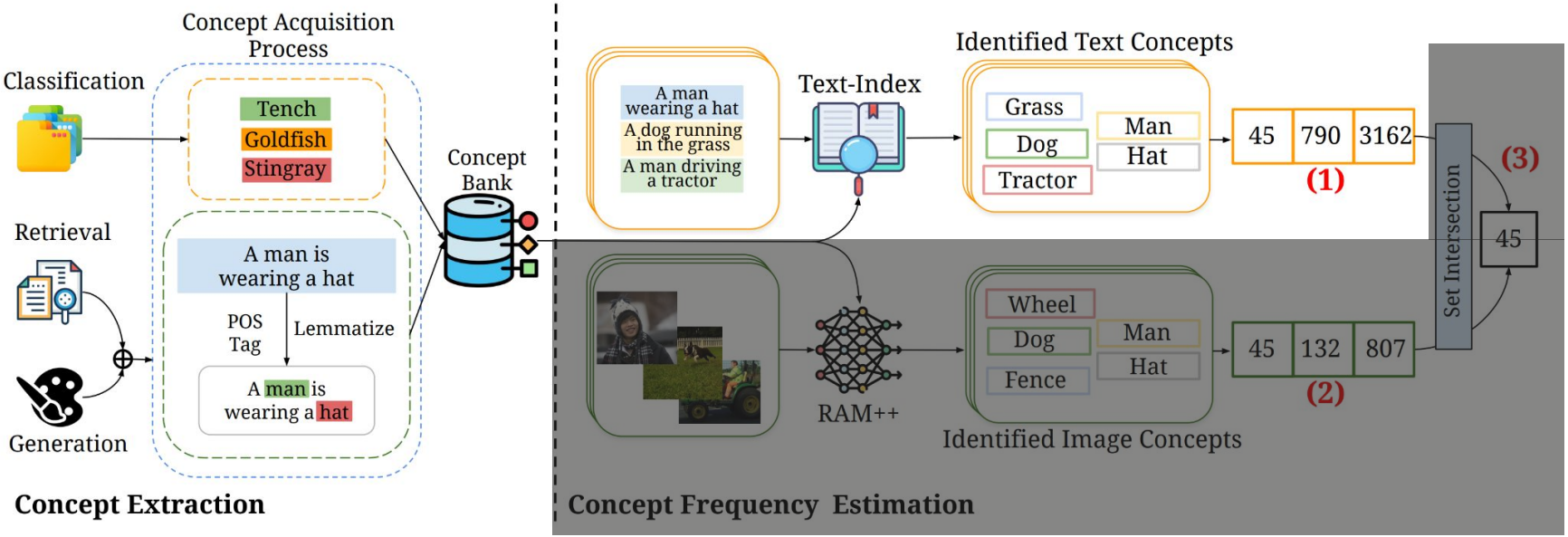
Characterising concept frequency

We first collate 4,029 concepts from 27 downstream tasks.



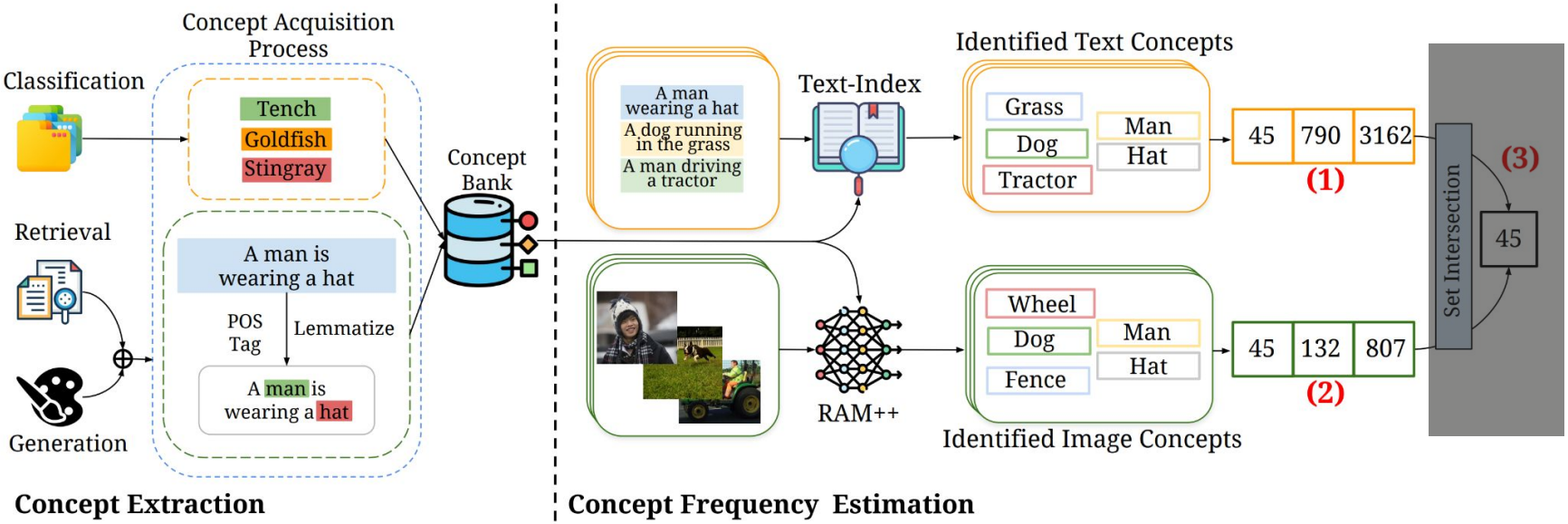
Characterising concept frequency

We then estimate concept frequencies in text captions of the pretraining datasets.



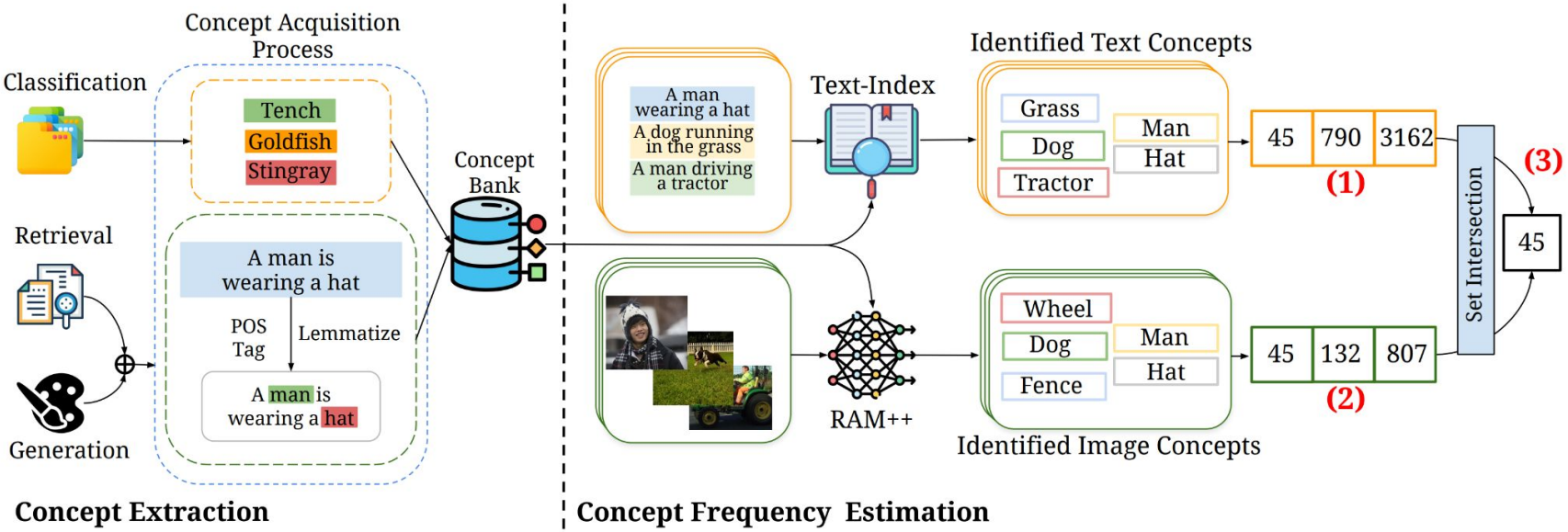
Characterising concept frequency

Next, we estimate concept frequencies in images of the pretraining datasets.

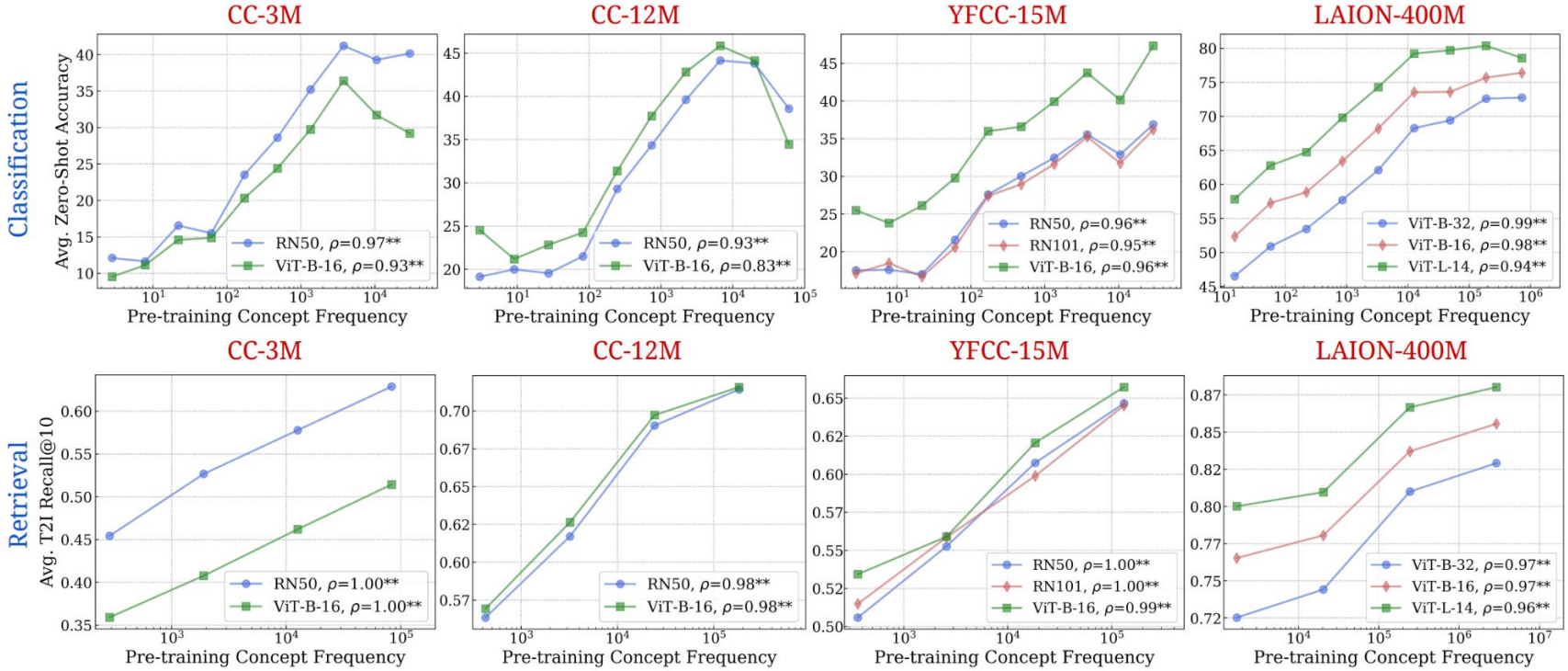


Characterising concept frequency

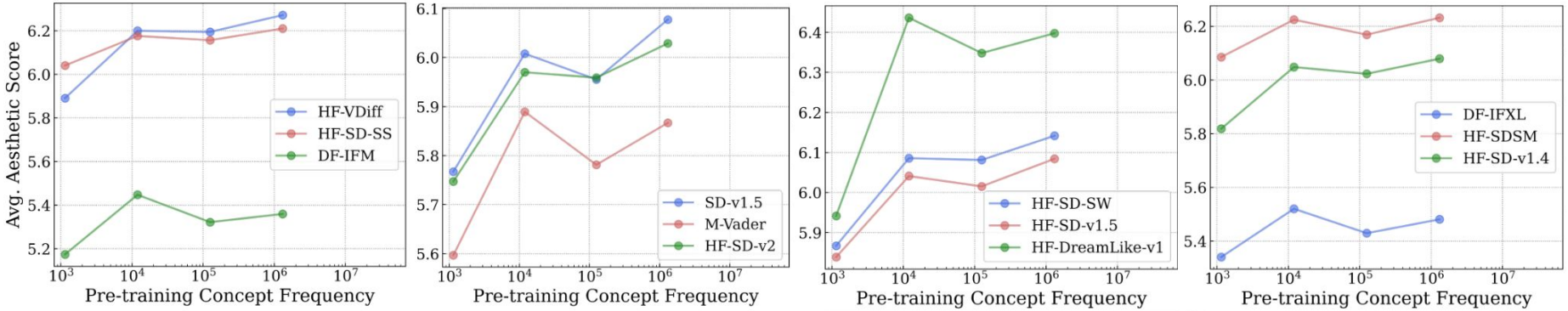
Finally, we estimate concept frequencies where both images and text captions capture the concept.



Key Result with CLIP models: Frequency determines Performance, log-linearly!



Key Result with text-to-image gen models: Frequency determines Performance, log-linearly!



Key Findings

- Log-linear scaling between concept frequency and zero-shot performance.
- To linearly improve performance, we have to scale up data exponentially!
- Extremely sample-inefficient (data-hungry) learning.

How robust is this result? Are there confounders?

We control for two important confounders:

- Similarity of pretraining samples to downstream samples

How robust is this result? Are there confounders?

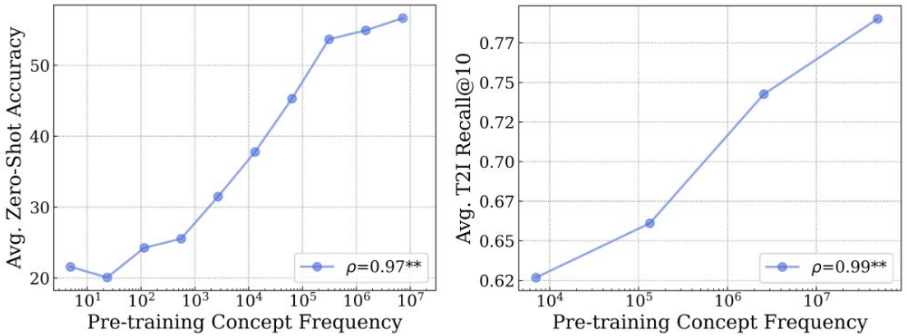
We control for two important confounders:

- Similarity of pretraining samples to downstream samples

- Synthetic and balanced pretraining data distribution

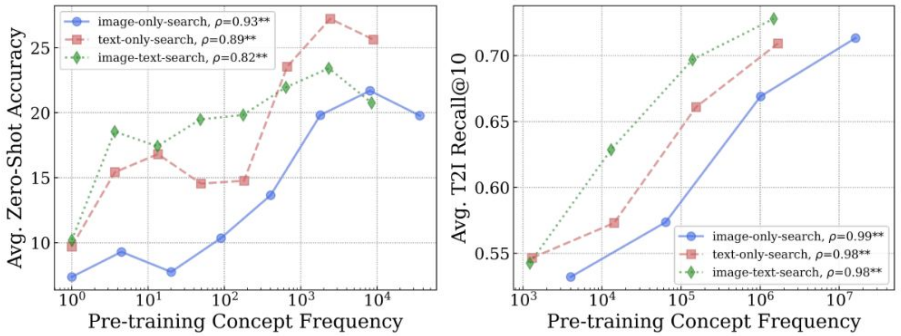
How robust is this result? Are there confounders?

Controlling for Similar Samples b/w Pretrain & Test Data



Classification Retrieval

Testing with Synthetic Pretraining Concept Distributions

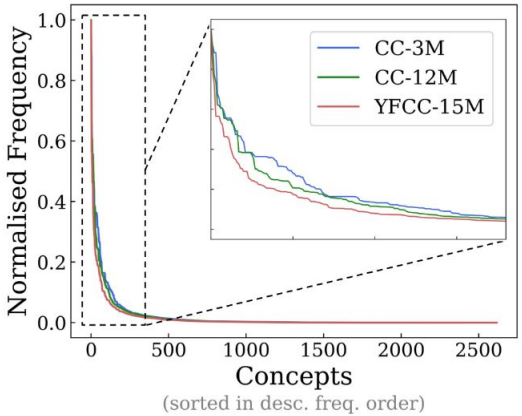


Classification Retrieval

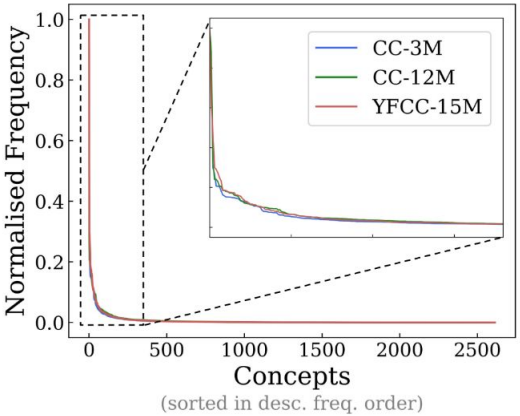
Additional data-centric nuggets: Insights for better data curation

- Pretraining datasets exhibit long-tailed concept distribution.
- Quantifying misalignment between concepts in image-text pairs.
- Concept frequencies across pretraining datasets are correlated

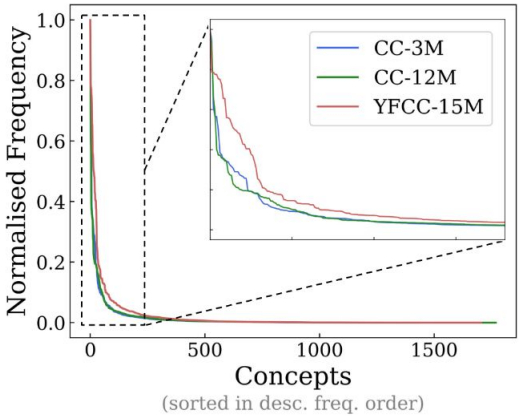
Pretraining datasets exhibit long-tailed concept distribution



(a) Text search counts

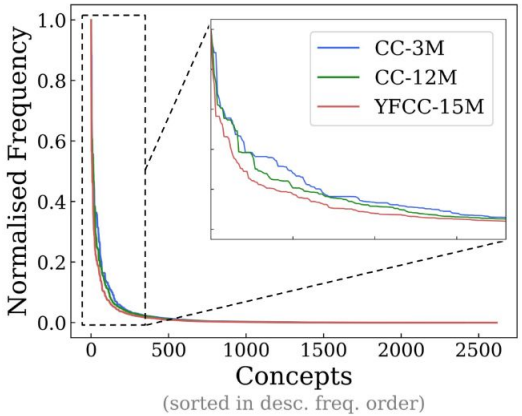


(b) Image search counts

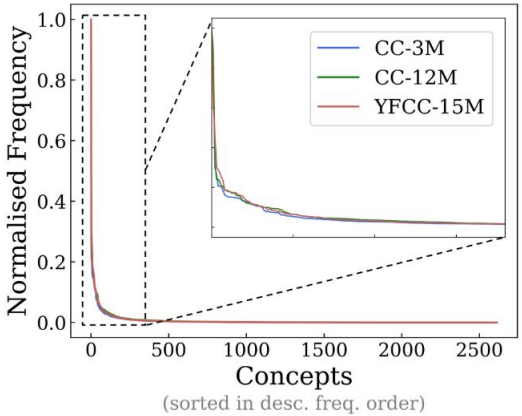


(c) Image-text search counts

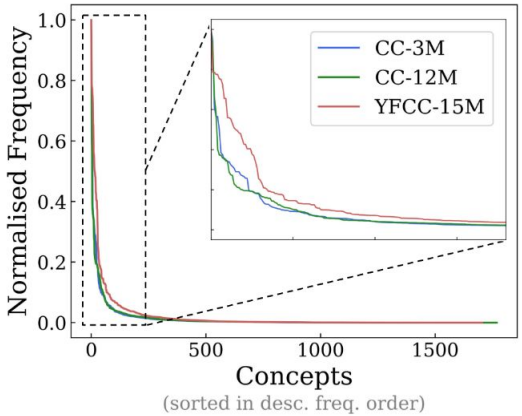
Pretraining datasets exhibit long-tailed concept distribution



(a) Text search counts



(b) Image search counts



(c) Image-text search counts

Even for explicitly balanced datasets (MetaCLIP, SynthCLIP). Why?

- Concepts are repeated. e.g., `sneaker`, `running_shoes`
- Concept co-occur. e.g., `woodpeckers` and `trees`

Quantifying misalignment between concepts in image-text pairs

Dataset/ Misalignment	Number of Misaligned pairs	Misalignment Degree (%)
CC3M	557,683	16.81%
CC12M	2,143,784	17.25%
YFCC15M	5,409,248	36.48%
LAION-A	23,104,076	14.34%
LAION400M	21,996,097	5.31%

Quantifying misalignment between concepts in image-text pairs

Dataset/ Misalignment	Number of Misaligned pairs	Misalignment Degree (%)
CC3M	557,683	16.81%
CC12M	2,143,784	17.25%
YFCC15M	5,409,248	36.48%
LAION-A	23,104,076	14.34%
LAION400M	21,996,097	5.31%

Reason why recaptioning methods are getting popular. Simple yet efficient way to improve “data quality”

Concept frequencies across pretraining datasets are correlated

Correlations	CC3M	CC12M	YFCC15M	L400M
CC3M	1.00	0.79	0.96	0.63
CC12M	–	1.00	0.97	0.74
YFCC15M	–	–	1.00	0.76
L400M	–	–	–	1.00

Concept frequencies across pretraining datasets are correlated

Correlations	CC3M	CC12M	YFCC15M	L400M
CC3M	1.00	0.79	0.96	0.63
CC12M	–	1.00	0.97	0.74
YFCC15M	–	–	1.00	0.76
L400M	–	–	–	1.00

All web-crawled data is ‘different’ yet ‘similar’. The web naturally induces a long-tailed distribution—best to make peace with it and find better curation and training strategies.

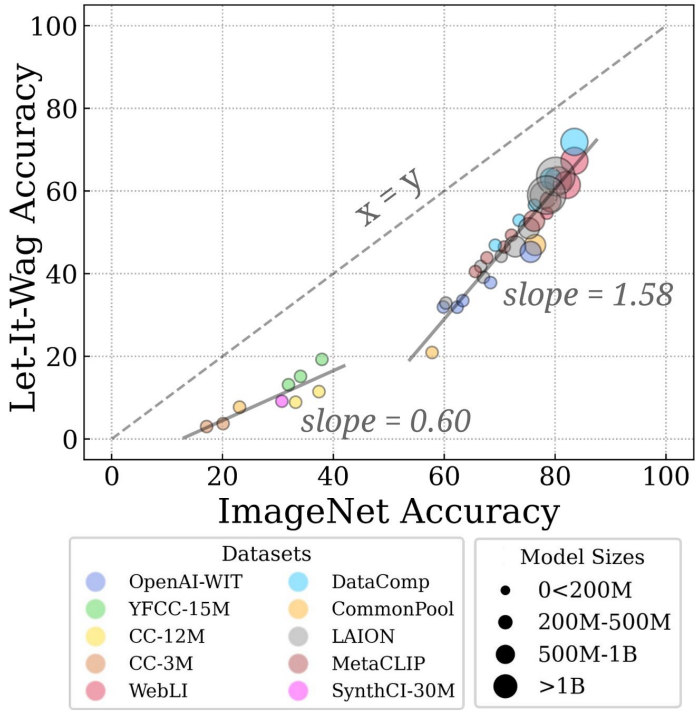
Testing the tail: *Let-It-Wag!*

To foster further research, we collect a true ‘long-tailed’ dataset.

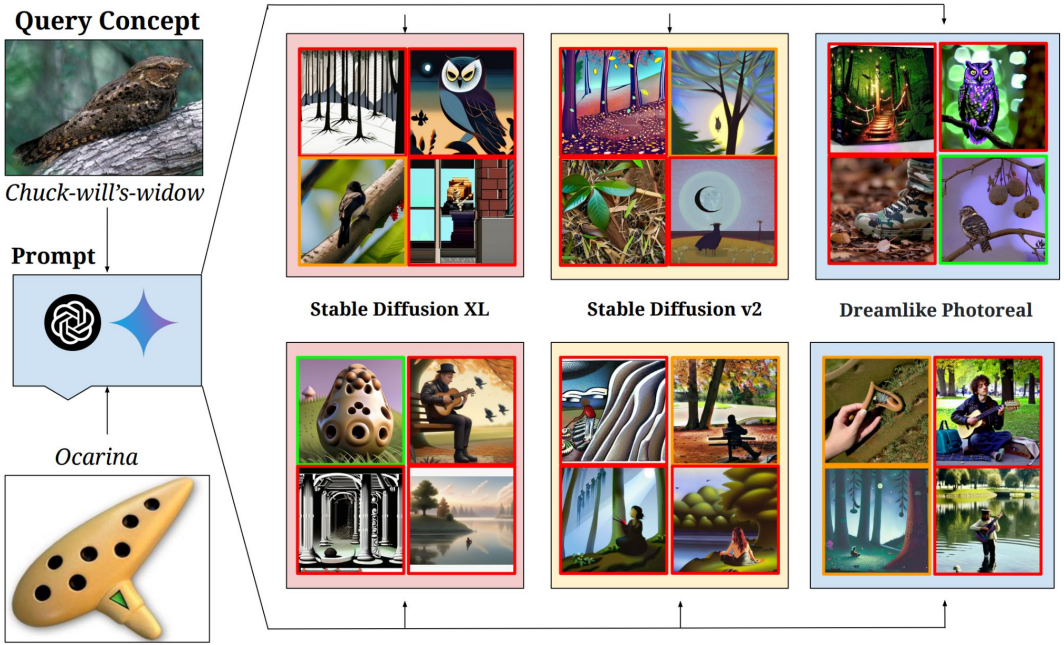
We curate images from 290 most infrequent concepts in LAION400M.

Test multiple image-text and text-to-image models.

Let-It-Wag! classification



Let-It-Wag! image generation



Summary and Insights for Data Curation

- The web is long-tailed and there is a lack of high-quality data for all use-cases
- Current multimodal models are extremely sample inefficient—to improve performance linearly, we need exponentially more data samples
- What do we want our models to generalize to?
- Downstream task-aware curation the way to go? Improving dataset priors improves model performance!

Open questions

- Effect of model scaling? Where do our results leave us with respect to scaling laws?
- Effects on compositional generalization?
- What are effective measures to curate data and combat the long-tailed nature?
  - Retrieval augmentation to the rescue?
  - Better balancing strategies while preserving diversity?
  - Is “quality filtering” always better?

Thanks for your attention!

Paper

Github

HuggingFace