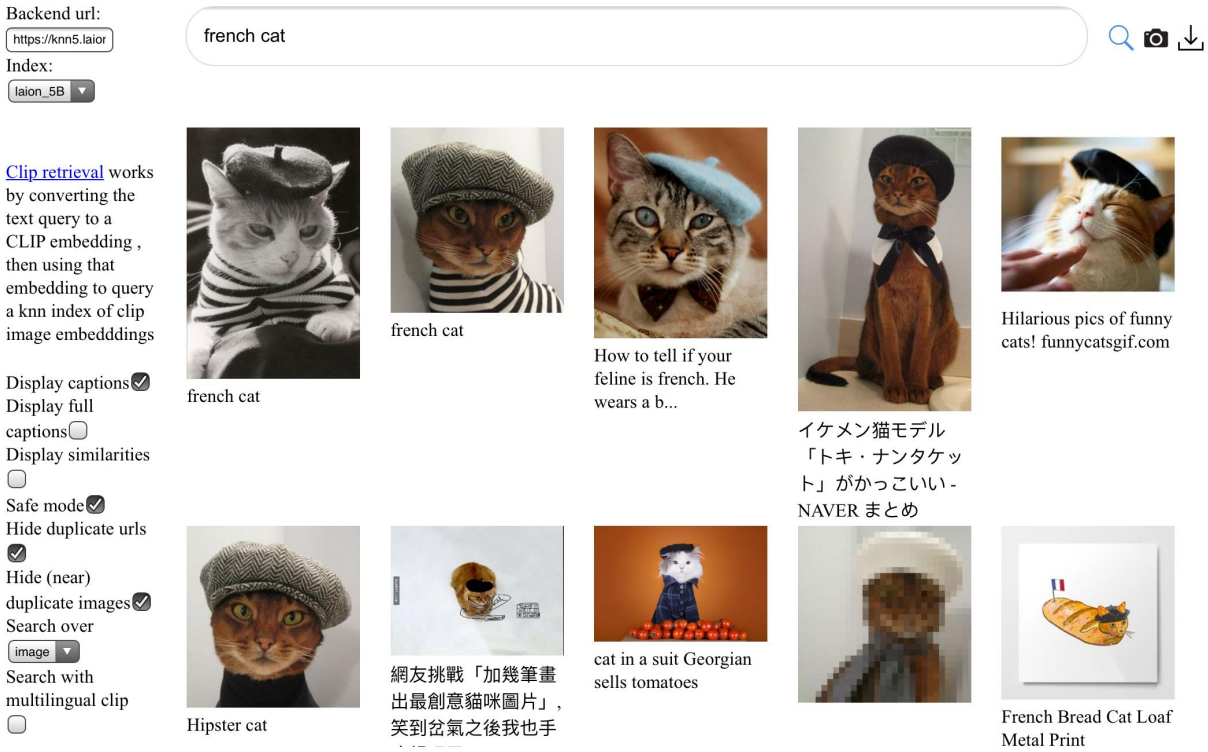
LAION-5B: A new era of open large-scale multi-modal datasets

We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world.

Authors **Christoph Schuhmann**, **Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev**



Large image-text models like ALIGN, BASIC, Turing Bletchly, FLORENCE & GLIDE have shown better and better performance compared to previous flagship models like CLIP and DALL-E. Most of them had been trained on billions of image-text pairs and unfortunately, no datasets of this size had been openly available until now.

To address this problem we present LAION 5B, a large-scale dataset for research purposes consisting of 5,85B CLIP-filtered image-text pairs. 2,3B contain English language, 2,2B samples from 100+ other languages and 1B samples have texts that do not allow a certain language assignment (e.g. names).

Additionally, we provide several nearest neighbor indices, an improved web interface for exploration & subset creation as well as detection scores for watermark and NSFW. We also announce a full reproduction of a clip training trained on LAION-400M at open_clip.

Explore the dataset at the search demo. See also the same post on laion website

We thank our sponsors huggingface, doodlebot and stability for providing us with computing resources to produce this dataset!

Disclaimer on dataset purpose and content warning

The motivation behind dataset creation is to democratize research and experimentation around large-scale multi-modal model training and handling of uncurated, large-scale datasets crawled from publically available internet. Our recommendation is therefore to use the dataset for **research purposes**.

Be aware that this large-scale dataset is uncurated. Keep in mind that the uncurated nature of the dataset means that collected links may lead to strongly discomforting and disturbing content for a human viewer. Therefore, please use the demo links with caution and at your own risk. It is possible to extract a “safe” subset by filtering out samples based on the safety tags (using a customized trained NSFW classifier that we built). While this strongly reduces the chance for encountering potentially harmful content when viewing, we cannot entirely exclude the possibility for harmful content being still present in safe mode, so that the warning holds also there.

We think that providing the dataset openly to broad research and other interested communities will allow for transparent investigation of benefits that come along with training large-scale models as well as pitfalls and dangers that may stay unreported or unnoticed when working with closed large datasets that remain restricted to a small community. Providing our dataset openly, we however **do not** recommend using it for creating ready-to-go industrial products, as the basic research about general properties and safety of such large-scale models, which we would like to encourage with this release, is still in progress.

Introduction

Since the release of CLIP & DALL-E in January 2021, several similar large multi-modal language-vision models have been trained by large groups. Models like FLORENCE, Turing Bletchley, ALIGN & BASIC demonstrated very strong transfer capabilities on novel datasets in absence of per-sample labels, which also steadily improved when growing training data amount, following scaling laws observed in previous research work.
These models require billions of image-text pairs to achieve competitive performances and unfortunately, no billion-scale image-text pair dataset had been openly available up until now.

To address this problem we release LAION 5B, a CLIP-filtered dataset of 5,85 billion high-quality image-text pairs, their CLIP ViT-L/14 embeddings, kNN-indices, a web interface for exploration & subset-creation and NSFW- and watermark-detection scores and tools.

We describe the procedure to create the dataset and demonstrate successful training of DALL-E architecture. Having sufficiently large scales, the dataset opens venues for research on multi-modal language-vision models to a broad community.

Download the data

We release the following packages under the LAION-5B project:

- laion2B-en 2.32 billion of these contain texts in the English language
- laion2B-multi 2.26 billion contain texts from 100+ other languages
- laion1B-nolang 1.27 billion have texts where a particular language couldn't be clearly detected.

The data can comfortably be downloaded with img2dataset

For training usage, we recommend reading the usage guide for training

In particular, we release this data:

- 5.85 billion pairs of image URLs and the corresponding metadata at laion2B-en laion2B-multi laion1B-nolang
- A knn index that enables quick search in the dataset
- Web demo of image-text search on LAION-5B clip-retrieval
- Safety tags at laion2B-en-safety laion2B-multi-safety laion1B-nolang-safety
- Watermark tags at laion2B-en-watermark laion2B-multi-watermark laion1B-nolang-watermark

The metadata files are parquet files that contain the following attributes: URL, TEXT, the cosine similarity score between the text and image embedding and height and width of the image. Watermark and safety tags can be joined with the metadata prior to downloading by using this script. Once that is done, they can easily be filtered upon with a probability threshold at your choice (we recommend 0.5 for safety and 0.8 for watermark).
You can also find the prejoined files at laion2B-en-joined laion2B-multi-joined laion1B-nolang-joined

License

We distribute the metadata dataset (the parquet files) under the Creative Common CC-BY 4.0 license, which poses no particular restriction. The images are under their copyright.

Dataset Statistics

We computed some statistics on the datasets to let people understand better:

Laion2B-en

Total: 2.3B samples
Number with height and width bigger than

- 256 -> 1324M
- 512 -> 488M
- 1024 -> 76M

Number of unsafe samples with a probability threshold of 0.5: 0.029
Number of watermarked samples with a probability threshold of 0.8: 0.061

Laion2B-multi

Total: 2.2B samples
Number with height and width bigger than

- 256 -> 1299M
- 512 -> 480M
- 1024 -> 57M

Top 10 languages:

	LANGUAGE	count	proportion
0	ru	241M	0.106
1	fr	168M	0.074
2	de	150M	0.066
3	es	149M	0.066
4	zh	143M	0.063
5	ja	131M	0.057
6	it	95M	0.042
7	pt	88M	0.038
8	nl	66M	0.029
9	pl	62M	0.027
10	no	49M	0.021

Number of unsafe samples with a probability threshold of 0.5: 0.033
Number of watermarked samples with a probability threshold of 0.8: 0.056

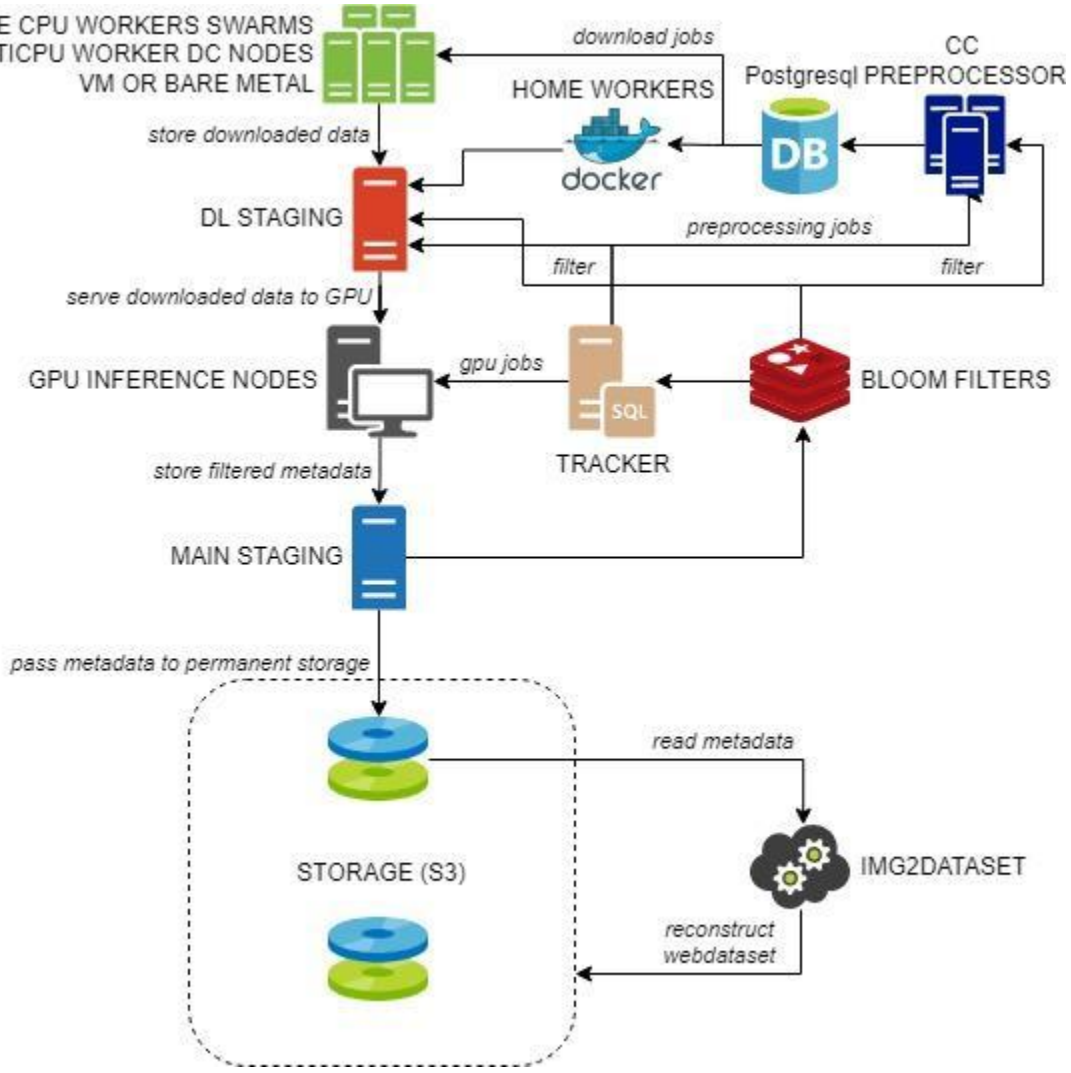
Laion1B-nolang

Total: 1.2B samples
Number with height and width bigger than

- 256 -> 1324M
- 512 -> 488M
- 1024 -> 76M

Number of unsafe samples with a probability threshold of 0.5: 0.03
Number of watermarked samples with a probability threshold of 0.8: 0.04

Acquisition pipeline



The acquisition pipeline follows the flowchart above and can be split into three major components:

- Distributed processing of petabyte-scale Common Crawl dataset, which produces a collection of matching URLs and captions (preprocessing phase)
- The distributed download of images based on shuffled data to pick a correct distribution of URLs, to avoid too heavy request loads on single websites
- Few GPU node post-processing of the data, which is much lighter and can be run in a few days, producing the final dataset.

Distributed processing of Common Crawl

To create image-text pairs, we parse through WAT files from Common Crawl and parse out all HTML IMG tags containing an alt-text attribute. At the same time, we perform a language detection on text with three possible outputs: English language with confidence, another language with confidence, no language which contains “no detection” and “detection under the confidence threshold”. The “no language” set often contains short texts, mostly with names of people and places.

All extracted information by the preprocessing workers were packed and sent to the Postgresql node for storage using the COPY command. The Postgresql server was maintained to keep about 500M records at all times by means of balancing the ingress and egress of data from the database.

Distributed downloading of the images

We download the raw images from the parsed URLs with asynchronous requests using Trio and Asks libraries in order to maximize all resources usage: vCPUs, RAM and bandwidth. We found that a single node in the cloud with 1-2 vCPUs, 0.5-1GB RAM and 5-10Mbps download bandwidth is inexpensive enough to allow downloading on a limited budget. Such a unit can process 10000 links in about 10-15 minutes. Each batch consisted of 10000 links taken from the Postgresql server by using the TABLESAMPLE technique, ensuring that the distribution among the 10000 links was following the distribution of the existing 500M records available on the database. We found that the distribution is still good when in the database are still above 20M records to be processed given that we had some 300 downloading workers at any time.

The above techniques allowed both maximizing downloading speed and minimizing IP reputation damages.

CLIP inference at the post-processing stage

The data pipeline continued with GPU nodes doing inference on the collected image-text pairs, and calculating the similarity of the embeddings for the image and the text. After the similarity score was established we removed the pairs under the threshold we decided to use, i.e 0.28 for the English dataset (with CLIP ViT B/32) and 0.26 for the rest (with mCLIP).

As an estimation, we removed about 90% of the samples, trimming the 50+ billion of candidates to just below 6 billion.

Filtering out unsuitable image-text pairs

After downloading the WAT files from Common Crawl, we apply the following filtering conditions:

- All samples with less than 5 characters alt-text length or less than 5 KB image size are dropped.
- All images with the too big resolution, potentially DOS bombs, were dropped before attempting to process them.
- Duplicate removal is performed with a bloom filter based on URL. Future runs would include more variate deduplication rules, such as URL + language for the multilanguage dataset.
- We use CLIP respectively MCLIP to compute embeddings of the image and alt-text. Then we compute the cosine similarity of both embeddings and drop all samples with cosine similarity below 0.28 for the English language (with CLIP B/32) and 0.26 for the multilingual dataset (MCLIP). These thresholds were selected based on human inspection of the test results.
- We use the CLIP embeddings of images and texts to filter out to the possible extent the illegal content.

Dataset preparation pipeline

After processing and filtering common crawl, 5,85B of URL/text samples remained.
We did additional steps after that in order to prepare the dataset.
See this semantic search blogpost and the readme of clip-retrieval for additional details about this process. See also semantic search at billions scale for more technical details of the process that was done for laion5B.

1. Downloading the data as webdataset with distributed img2dataset
2. Computing Vit-L/14 embeddings with distributed clip-inference
3. Computing a KNN index from these embeddings using autofaiss
4. Computing additional tags (NSFW and watermark) using clip embeddings

Distributed img2dataset

We developed the img2dataset library to comfortably download from a given set of URLs, resize and store the images and captions in the webdataset format. This allows downloading 100 million images from our list of URLs in 20 hours with a single node (1Gbps connection speed, 32GB of RAM, an i7 CPU with 16 cores), which allows anyone to obtain the whole dataset or a smaller subset.

For LAION-5B we introduced a distributed mode for this tool, allowing to downloading the 5,85B samples in a week using 10 nodes.

Distributed clip inference

From these images, the clip retrieval inference tool was used to compute ViT-L/14 embeddings, allowing for a better analysis capacity of the data. In particular, a distributed mode made it possible to compute these embeddings in a week using 32 A100: this larger clip model can only be computed at a speed of 312 sample/s per GPU, compared to 1800 sample/s for ViT-B/32.

The resulting embeddings are available for everyone to use e.g. for clustering, indexing, linear inference.

Distributed indexing

We then used these 9 TB of image embeddings to build a large PQ128 knn index using the autofaiss tool. To make this run faster, a distributed mode is available.

Integration in the search UI

In order to demonstrate the value of this data, we integrated this index into the knn search UI. It is powered by the code called clip back.
The knn index is 800GB and the metadata (URL and captions) as well, so memory mapping is used for both in order to use no ram, only an SSD drive of that capacity is required.

Watermark and safety inference

We wanted to give users the ability to remove unsafe examples, and watermarked examples. To do that we collected training and test sets. The training set was augmented with examples retrieved from the knn index, while the test set samples were selected to represent well the dataset distribution, but were all manually annotated.

The inference is done using the embedding-reader module for NSFW and LAION-5B-WatermarkDetection for watermarks

These tags were also integrated into the UI, allowing everyone to observe that the safety tags indeed filter out almost all the unsafe results, and giving confidence that training a generative model on this data will not result in unexpectedly unsafe images.

Watermark

The training dataset is 90000 samples (45222 watermarks, 44778 clear).

Watermarked images are a big problem when training generative models like DALL-E or GLIDE. To tackle this problem we trained a watermark detection model and used it to calculate confidence scores for every image in LAION-5B.

Therefore we created a training dataset consiting of 90.000 images with 50% watermarked and 50% clean images.
The majority of the watermarked images have been extracted from the LAION-400M kNN index through the use of several text prompts like “clip art watermark”, “cat watermark” or “landscape watermark”.

The images in the cleaned category were composed of images from the Open Images dataset and images that contained texts, but no watermarks, like PPT slides and memes, also retrieved from the kNN indices of LAION-400M.

While we tried to curate a test set to evaluate the quality of our watermark detection model, we realized that it is almost impossible to draw a clear line between what actually is a watermark and what is not. For example pictures with small transparent texts at the bottom had been considered by some people as watermarked, by others not.

In the end we decided to choose a model based on our consensual judgment. It seems to be “good” at spotting obvious watermarks like those used on popular stock image sites.

The creation of high-quality, openly accessible watermark detection test sets with clear and plausible definitions of what should be considered a watermark and what not, remains a challenge for future projects.

Nevertheless we are convinced that removing images with a high confidence score for containing a watermark based on our model will significantly reduce the percentage of images that would be considered as obvious watermarks.

The model is available at https://github.com/LAION-AI/watermark-detection and https://github.com/LAION-AI/LAION-5B-WatermarkDetection/releases/tag/1.0

Safety

On a balanced manually annotated safety test set with 3000 samples:

- the accuracy of the B32 NSFW classifier is: 0.960
- the accuracy of the ViT L 14 NSFW classifier is: 0.961

The model, as well as the training code, are available at CLIP-based-NSFW-Detector
The tags are available at laion2B-en-safety laion2B-multi-safety laion1B-nolang-safety

Demo at clip-retrieval (check/uncheck safe mode)

Using laion datasets

Laion5B and LAION-400M can be used to train

- Generative models: training image/text generative models, e.g autoregressive models like DALL-E or diffusion models like GLIDE
- Models with contrastive losses: self-supervised training on image/text pairs using contrastive losses, e.g CLIP
- Classification models: e.g, performing zero-shot classification by extracting pseudo labels from queries on the dataset

We present here a few examples of models that were trained on laion datasets with success.

CLIP

We, LAION, are currently working together with the Cross Sectional Team Deep Learning (CST-DL), Scalable Learning and Multi-Purpose AI Lab (SLAMPAI) at the Jülich Supercomputing Centre (JSC) and the Open CLIP team in the replication of OpenAI's CLIP results.

At the time of writing, we just finished the training of a CLIP ViT-B/32 on LAION-400M that matches the performance of OpenAI's original ViT-B/32 CLIP.

Dataset	CLIP - Accuracy of ViT-B/32 (OpenAI)	CLIP - Accuracy of ViT-B/32 (Cade)
ImageNet	63.2	62.9
ImageNetV2	-	62.6
Birdsnap	37.8	46.0
Country211	17.8	14.8
Flowers102	66.7	66.0
GTSRB	32.2	42.0
Stanford Cars	59.4	79.3
UCF101	64.5	63.1

(The results in the right column are from our model. - huge thanks to Cade Gordon & Ross Wightman for performing the training run)

The repository with the training code and the model checkpoints can be found here:

https://github.com/mlfoundations/open_clip

We gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this part of work by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS Booster at Jülich Supercomputing Centre (JSC).

BLIP inference tuning

BLIP is a model that was trained for both image-text matching and image captioning.

It was trained on a 115M subset of LAION-400M.

To improve the results of the generated captions we (LAION) performed over 100 experiments to determine the hyperparameters that maximize the BLEU-4 score compared to MS COCO captions.

Here you can see some of our results.

eval_best_auto0185: An orange cat is looking at its reflection in the mirror.

eval_best_auto0190: A green highway sign with the words Queens Bronx.

We found that we can significantly improve the quality of the captions by generating 40 (or more) candidate captions for each image and then ranking them using OpenAI's CLIP ViT-L/14 & CLIP-Resnet50x64.
First we ranked all candidates with ViT-L/14 and then we ranked the top-5 results again using Resnet50x64.
Preliminary results of human evaluations indicate that:

1. our evaluators gave the generated captions an average quality rating of 3,8 on a scale from 0 to 5, with a standard deviation of 0,9 (in this paricular hyperparameter configuration n= 600)
2. our evaluators gave original human captions from MS COCO an average quality rating of 3,9 with a standard deviation of 0,8 (n = 2100)

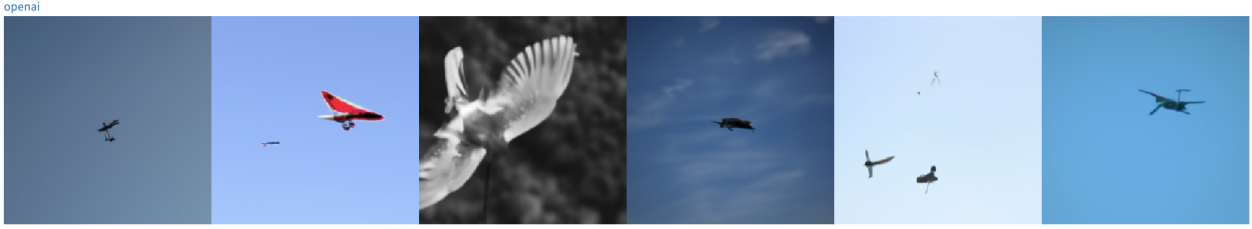
—> We hypothesize that the generated captions match (& sometimes even surpass) the average quality of the human captions of MS COCO (which are sometimes also far from perfect) in most cases, but sometimes (in less than <10%) contain obvious mistakes, that humans would not make, because deeper kind of world knowledge & „common sense" would be necessary in those cases.

GLIDE

Clay Mullis (alias afiaka87) used LAON-2B to fine-tune the OpenAi glide model and managed to reintroduce human generations.
Samples

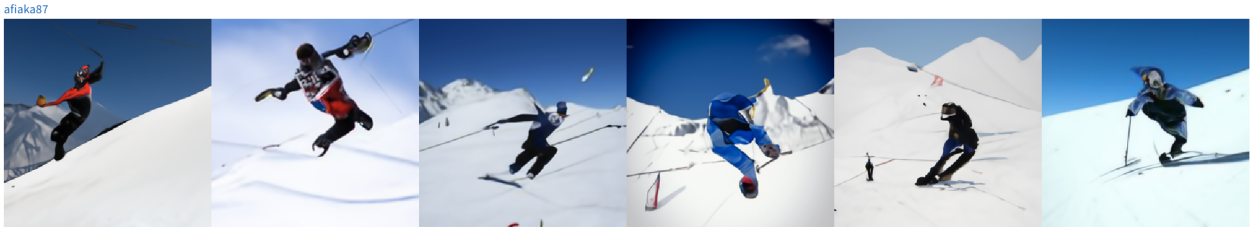
- https://replicate.com/afiaka87/laionide-v4
- https://wandb.ai/afiaka87/glide_compare/reports/Finetuning-GLIDE-on-Laion5B--VmlldzoxNTg3MTkz
- https://wandb.ai/afiaka87/laionide-v3-glide/reports/Laionide-Version-3-Benchmark--VmlldzoxNjE0MTE3

openai



A person on skis flies high in the air, while someone down below takes a picture.

afiaka87



A person on skis flies high in the air, while someone down below takes a picture.

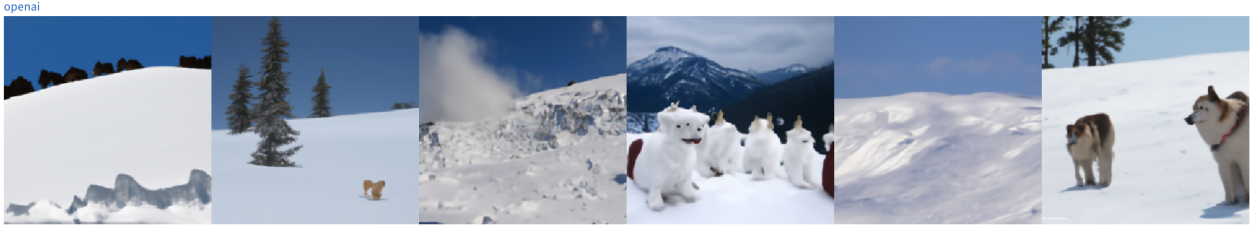
openai

a group of people sitting at a table eating supper

afiaka87

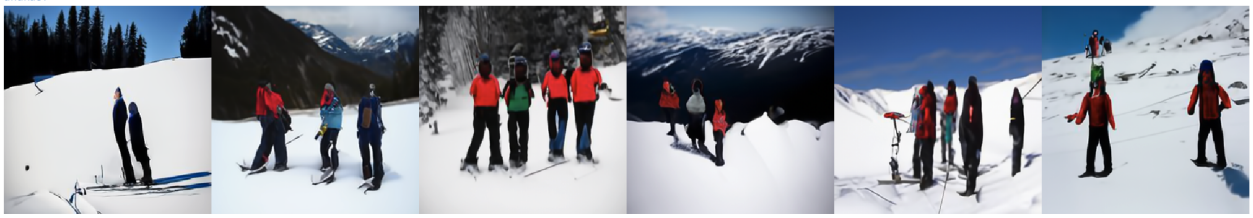
a group of people sitting at a table eating supper

openai



three people standing next to each other wearing skis and standing on a snow covered slope

afiaka87



three people standing next to each other wearing skis and standing on a snow covered slope

Semantic search and subset extraction

The clip-retrieval interface allows a user to search images and texts based on a query image or text using the CLIP embeddings of the input and our precomputed kNN indices. It demonstrates the diversity of images and captions that can be found in LAION-5B as well as high semantic relevance shows the distribution of image sizes of LAION-5B. Given the abundance of high-resolution images, one can produce subsets of images for training various customized models, and also choose image resolution that is suitable for the purpose of particular training.

CLOOB

Katherine Crowson and John David Pressman recently trained a CLOOB ViT-B/16, variant of CLIP, for 32 epochs on LAION-400M and got preliminary results, that come close to the performance of OpenAI's ViT-B/32, even though this was an early run with unoptimized hyperparameters.
The checkpoints can be found here:
https://github.com/crowsonkb/cloob-training

Model name	Top 1	Top 5
cloob_laion_400m_vit_b_16_16_epochs	0.61238	0.8492
cloob_laion_400m_vit_b_16_32_epochs	0.62816	0.85964
OpenAI CLIP ViT-B/32	0.6327	0.88772
OpenAI CLIP ViT-B/16	0.68132	0.91768
OpenAI CLIP ViT-L/14	0.75388	0.9454
OpenAI CLIP RN50	0.59806	0.86498
OpenAI CLIP RN101	0.62296	0.88106
OpenAI CLIP RN50x4	0.66268	0.9046
OpenAI CLIP RN50x16	0.70754	0.92822
OpenAI CLIP RN50x64	0.74134	0.94146

zero-shot accuracies on Imagenet-1K

We are in touch with Andreas Fürst, one of the original CLOOB authors, and learned from him that their team is currently (at the time of writing) training a CLOOB ViT-B/32 with LAION-400M with optimized hyperparameters and very promising results so far (53% zero-shot accuracy on Imagenet after 7 epochs).

Papers citing LAION 400M

After the release of LAION-400M, several papers used LAION-400M for image generation, text to image generation, image to text generation and text image matching:

- Vector Quantized Diffusion Model for Text-to-Image Synthesis used LAION-400M to train VQ diffusion text to image generation models
- High-Resolution Image Synthesis with Latent Diffusion Models used a subset of LAION-400M to train latent diffusion models
- General Facial Representation Learning in a Visual-Linguistic Manner LAION-400M face subset to train a face clip
- BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation image captioning using LAION-400M subset
- MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning was trained on image question answering using a LAION-400M subset

Conclusion

By releasing an updated version of an openly available dataset that contains 5 billion image-text pairs, we have set new Standards for the scale of openly available datasets and enable researchers from all over the world to train state-of-the-art language-vision models like GLIDE or Turing Bletchley.

As proof of concept, we demonstrated that a subset of our dataset can be used to train various CLIP-like models, producing samples of sufficient quality. This dataset extends the possibilities in multi-language large-scale training and research of language-vision models, that were previously restricted to those having access to proprietary large datasets, to the broad community.

What's next?

This is only the beginning! Now that this huge and open dataset is released, it can be used to train many models, such as gigantic clip models, image/text generation models and much more.

We have so many projects going on that it's probably best, if you are interested, to join our Discord server and check out what's going on.

We are and always will be a grassroots community that works openly and welcomes everyone who is kind and passionate and for machine learning.

Join us in discord and help us to train models like CLIP, BLIP, GLIDE, Dall-E, SimMIM, AudioCLIP and don't hesitate to share your ideas for new projects with us.

Become a part of our constantly growing crowd of supporters who help us to make machine learning dreams come true!

Credit Assignment

- **Christoph Schuhmann**: He led this project and built POCs for most of its components including clip filtering,the safety model, the watermark model and the Blip inference tuning project.
- **Richard Vencu**: System architecture and download script optimizations, GPU assisted filtering. Set up the AWS infrastructure.

- **Romain Beaumont**: Guidance on scaling for the common crawl filtering pipeline. Built and ran the dataset preparation pipeline: pyspark deduplication job, img2dataset, clip inference, autofaiss, safety tags.
- **Clayton Mullis**: DALLE-pytorch training/analysis, glide training, WDS filtering
- **Jenia Jitsev**: scientific organization & writing, experiments planning and design, compute resource acquisition, general supervision
- **Robert Kaczmarczyk**: Established WDS architecture, performed DALL-E training runs, balancing calculation, sample (NSFW, watermark, caption quality) annotation and manuscript revision
- **Andreas Köpf:** He conducted the hyperparameter search for the inference strategies with the BLIP image-captioning model
- **Aarush Katta:** Trained the watermark model
- **Cade Gordon**: Run distributed inference for the watermark tags & trained the CLIP B/32 model on JUWELS Booster
- **Ross Wightman:** Ross helped Cade with the debugging & training of the CLIP-B/32 model and executed experiments on JUWELS Booster
- **Katherine Crowson and John David Pressman:** Trained the CLOOB model
- **Aran Komatsuzaki**: Led an image-text-pair dataset building project, which inspired this project.
- **Bokai Yu**: Accomplished most of the work to make the knn index building tool autofaiss work in a distributed setting