Supplementary: Reproducible scaling laws for contrastive language-image learning

A. Further details on distributed training

A.1. Supercomputer specifications

The JUWELS Booster [34] supercomputer used for training consist of 936 compute nodes that host four NVIDIA A100 GPUs each, providing 3744 GPUs in total. The installed A100 Tensor Core GPUs (40 GB) provide 19.5 TFLOP/s of FP64$_{\text{TC}}$ computing performance each. The GPUs are hosted by AMD EPYC 7402 CPUs with 2×24 cores (SMT-2) per node, clocked with 2.8 GHz. Each node is diskless and is equipped with 512 GB of RAM. The network is based on Mellanox HDR200 InfiniBand, with four Mellanox ConnectX 6 devices per node, each providing 200 Gbit/s bandwidth per direction.

The NVIDIA A100 GPUs reach peak efficiency of 48.75 GFLOP/(s W) when utilizing the FP64 Tensor Cores. This made the employed machine rank highest in the Green500 list as of November 2020 as the most energy efficient supercomputer among the first 100 machines of the Top500 list with 25 GFLOP/(s W).

A.2. Scaling and training time

Here, we report scaling behavior during large-scale pre-training using ViT-L/14 as a vision backbone with OpenCLIP [32]. We performed scaling experiments to assess the scalability of data parallel training distributed across many GPUs on multiple nodes using PyTorch DDP. The efficiency in Figure 6b is computed using the following formula: $E(N) = 100 \times \frac{T(N)}{N \times T(1)}$. $T(N)$ is the total measured throughput in Im/s for N GPUs. The best achievable efficiency, when scaling is perfect, is 100%.We observe that scaling is sufficiently close to ideal linear, staying above $\approx 84\%$ for 1024 GPUs (256 nodes). We also provide the raw throughput (Im/s) numbers in Figure 6a.

A.3. Sharding contrastive loss

The InfoNCE loss [52] used by CLIP can be thought of as a method to maximize the mutual information between text and image representations. Formally, Oord et al. express that $I(X;Y) \geq \log(N) - \mathcal{L}_N$, N denoting batch size and $\mathcal{L}_N$ representing the InfoNCE loss. As a result of this lower bound, maximizing the batch size will maximize our mutual information.

Radford et al. [55] take advantage of this bound and use $N = 32,768$ to train CLIP. Such a batch size necessitates the sharding of computation. Although the original CLIP paper points towards this notion, the implementation details are nontrivial.

Before sharding, the similarity scores will take up $\mathcal{O}(N^2)$ memory on each worker, totalling to 4 GB of VRAM in FP32. After sharding memory reduces to instantiating two $n \times N$ matrices, n being the batch size allocated to each worker. Using a local batch size of 256, the similarity matrices now occupy 64 MB of memory in FP32.

To achieve this memory reduction, we can eliminate redundant computations and compute the similarities of local features versus all features. When aggregated across all machines, this achieves identical gradients. However, it should be noted that the all-gather method is imperative for correct gradient calculation. PyTorch's standard `torch.distributed.all_gather` can not be differentiated through, while `torch.distributed.nn.functional.all_gather` can be. Thus, we require the use of the latter to correctly calculate the gradients in a distributed manner.

A.4. Training instabilities

As parameterization increased within our training runs, so did model model instability. Half-way through the runs of ViT L/14 H/14 and g/14, NaN values and loss spikes began occurring.

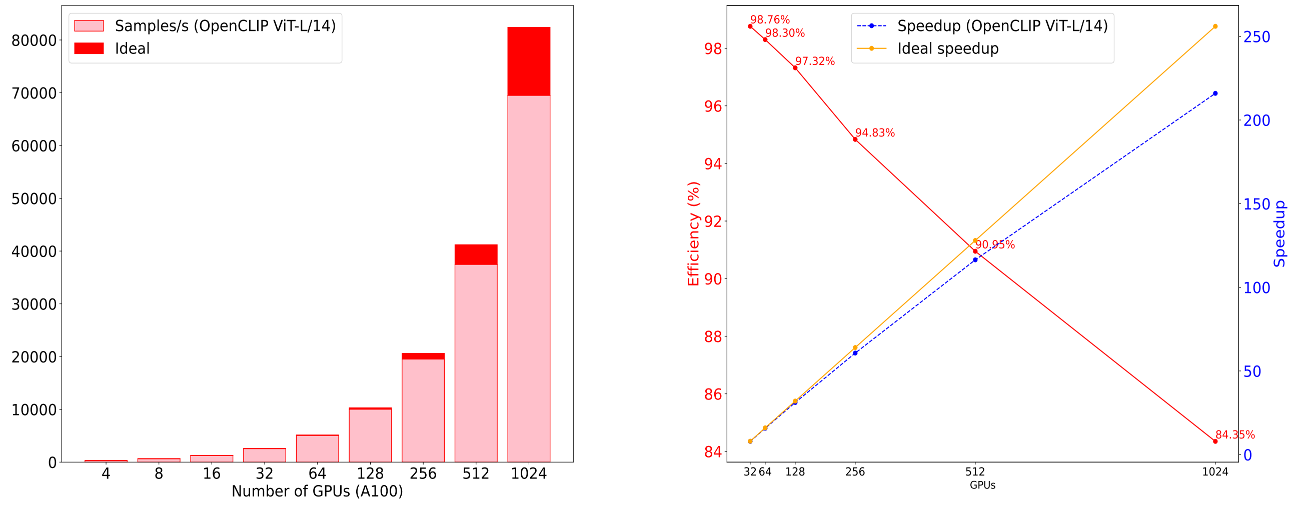
To address these issues, we attempted to use extra normalization layers, add scaled cosine attention, resume many steps before crashes, and implement other architecture tweaks with no success. What ended up solving the stability issues was increasing precision.

Using Automatic Mixed Precision (AMP) with bfloat16 over float16, or float32 with tensor-float32 resolved the issues mentioned above. We also have observed that even the smaller ViT-B models with AMP can become unstable when learning rate and batch size become sufficiently large, suggesting a generic scheme behind the phenomenon where frequency of instabilities occurring during the training is a function of model scale and global batch size.

B. Experimental details

B.1. Datasets employed in experiments.

LAION-400M and LAION-5B. Both LAION-400M [66] and LAION-5B [65] are open, public image-text datasets that were composed by obtaining links from Common Crawl [1]. While LAION-400M contains 414M english image-text pairs, LAION-5B is currently the largest public image-text dataset containing over 5.8 billion multilingual image-text examples. In both cases, samples are obtained by filtering a subset of Common Crawl with a pre-trained OpenAI ViT B/32 model. LAION-5B contains an English image-text subset of 2.32 billion samples, to which we refer as LAION-2B in this work. Besides the open nature of the datasets, a further advantage is full transparency about the dataset composition and assembly, with software



(a) Throughput in Im/s

(b) Speedup and efficiency

Figure 6. Distributed training for OpenCLIP ViT-L/14, scaling behavior on the supercomputer using A100 GPUs while varying the number of GPUs. In Figure 6a, we show the raw throughputs and in Figure 6b we show speedup and efficiency we obtain in the same setup, relative to training with a single node (each node contains 4 GPUs).

Dataset	# English Img-Txt Pairs
Public Datasets	
LAION-400M	407M
LAION-2B	2.3B
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

Table 2. **Open LAION datasets used for pre-training in this study.** Adapted from [65]. LAION-2B is a subset of multi-lingual LAION-5B and is more than 20 times larger than other public English image-text datasets. The scale of LAION-2B is comparable to the largest private dataset used for language-vision model training.

stack and tools around LAION-400M and LAION-5B released as open-source, increasing reproducibility of experiments. This already resulted in numerous works using the datasets for training state-of-the-art language-vision models [19, 27, 49, 63, 76], validating the usage of those datasets for studying scaling laws in this work.

Downstream transfer and fine-tuning datasets. For downstream classification tasks, in addition to standard ImageNet, we follow [65] and use VTAB+, a collection of datasets in VTAB together with ImageNet derived robustness datasets and additional datasets, forming a comprehensive set of 35 tasks. For evaluating retrieval, we make use of MS-COCO and Flickr30K. For fine-tuning, we make use of a dedicated ImageNet-12k dataset (12M training examples, 470K validation examples) which is a subset of the full ImageNet-22k (14M examples) that we employ for the multi-stage fine tuning procedure described in Sec. 4.4. For more details on downstream datasets, refer to Table 26.

Duplication check for pre-training and downstream datasets.. To ensure that images from downstream datasets are not contained in LAION, we conduct a simple duplication check based on the perceptual image hash library pHash [82]. We apply pHash's discrete cosine transform (DCT) method on LAION-400M images and images from downstream datasets. Afterwards, for each downstream dataset, we count the number of duplicates by finding the hashes that are also present in LAION-400M. We provide the overlap percentage found on a subset of downstream datasets in Table 3. In Figure 7, we also provide a sample of images from downstream datasets detected as duplicates in LAION-400M. Overall, the ratio of detected duplicates is around 1%, except on ImageNet-R (3.80%) and ImageNet-Sketch (5.15%). We investigate further and re-evaluate zero-shot performance of our pre-trained Vit-H/14 on ImageNet-R and ImageNet-Sketch by removing duplicates from their test sets. For ImageNet-R, zero-shot top-1 accuracy goes from 89.32% to 89.21% after removing duplicates. For ImageNet-Sketch, zero-shot top-1 accuracy goes from 66.57% to 66.59% after removing duplicates. We conclude, based on those results, that it is unlikely that downstream results would be affected by the duplicates. This would be in line with previous works [55, 86] which explicitly measured and compared

Dataset	Overlap%
ImageNet	1.02
ImageNet-v2	1.35
ImageNet-R	3.80
ImageNet Sketch	5.15
ImageNet-A	0.40
ObjectNet	0.10
CIFAR-100	0.02
CIFAR-10	0.03
MS-COCO	1.12
Flickr30K	1.30

Table 3. Ratio of images (%) on downstream datasets that were detected on LAION-400M, using pHash [82].

Model	ImageNet top-1 (%)	MS-COCO Recall@5 (%)
H/14 (68B)	79.73	75.03
g/14 (34B)	79.11	74.48
g/14 (68B)	80.66	75.85
G/14 (13B)	78.26	73.75
G/14 (34B)	80.47	75.68
G/14 (68B)	81.92	76.99

Table 4. Performance extrapolation of g/14, H/14 and G/14 on larger scales corresponding to Fig.8. We fit a power-law on the Pareto frontier of available models. We show the zero-shot top-1 accuracy predictions for ImageNet and zero-shot retrieval image retrieval Recall@5 predictions for MS-COCO.

performance on deduplicated downstream datasets, reporting that duplication in test sets do not significantly alter most results. This is likely due to the very large scale and diversity of pre-training data. We leave more elaborated duplication detection procedures for future work.

B.2. Further experimental results

B.2.1 Predictions derived from scaling laws

We can use scaling laws derived from our measurements to predict model performance for larger scales on different downstream tasks. To perform predictions, we fit a power-law on the Pareto frontier[3]. Fig.8a and Fig.8b show extrapolation of performance for ImageNet and MS-COCO, respectively. According to the predictions, H/14 (68B samples seen) would achieve 79.73% (+1.76%) zero-shot top-1 accuracy on ImageNet and 75.10% (+1.60%) image retrieval Recall@5 on MS-COCO, compared to our trained H/14 (34B samples seen). See also Tab.4 for more detailed extrapolation numbers. For g/14 (68B samples seen), we predict 80.66% (+4%) zero-shot top-1 accuracy on ImageNet and 75.85% (+3.45%) image retrieval Recall@5 on MS-COCO, compared to our trained g/14 (13B samples seen). On the largest compute budget we consider, G/14 (68B samples seen), we predict 81.92% zero-shot top-1 accuracy on ImageNet and 76.99% image retrieval Recall@5 on MS-COCO.

B.2.2 Fine-tuning

In Table 9, we show detailed results of fine-tuning on ImageNet with and without extra data (Imagenet-12k), and show results of the fine-tuned models on five ImageNet robustness test sets. Also, complementing the results shown in Figure 5 in Section 4.4, we show a per-task breakdown of the the zero-shot and fine-tuned performance on the eight classification tasks in Figures 9 and 10. Exact numbers are shown in Tables 6, 7, and 8.

Moreover, since fine-tuning on some downstream tasks can decrease accuracy on others, we experiment with model patching by interpolating between the weights of fine-tuned and zero-shot models, as in Ilharco et al. [31].[4] We choose the mixing coefficient $\alpha \in 0, 0.1, ..., 1.0$ that maximizes average accuracy on the eight downstream tasks, while accuracy on ImageNet—used as a control—decreases by one percentage point or less. In Figure 11, we show how scale affects performance on the eight tasks we fine-tune one, along with that on ImageNet.

Finally, Tables 10 and 11 include hparam templates for reproducing ImageNet fine-tune results. Once published, the individual model weights will include their specific training hyper-parameters as there is some variation in specific instances (i.e. at different upscale sizes, from 12k to 1k). Motivated by BEiT [4], all ImageNet fine-tune runs make use of layer-wise learning-rate decay (also known as discriminative fine-tuning [29]); this is an important parameter that needs tuning per model size along with the learning-rate itself.

B.2.3 Control experiments

Batch size during pre-training. To be able to train efficiently on a large number of GPUs (up to 1520 in this work), it is desired to maximize the local batch size for each GPU worker for performing data parallel distributed training. For this large amount of GPUs, it leads to training with global batch sizes of 86K-88K. As we would like to also re-use experiments that were already performed with smaller batch sizes of 32K-45K, we execute control experiments to reassure that varying batch size in those ranges does not alter observed model performance on downstream tasks strongly. The experiments summarized in Table 12 provide evidence that performance variation due to changes in batch size is

[3] Since total compute budget (measured in GMAC) of different trained models are not exactly aligned, we adopt a binning approach. We bin the GMAC compute budget axis and compute the optimal performance within each bin, then fit a line in log-log space on the resulting bins.

[4] The weights θ_{patched} of the patched model are obtained via the equation $\theta_{\text{patched}} = (1-\alpha)\theta_{\text{zero-shot}} + \alpha\theta_{\text{fine-tuned}}$, where $\alpha \in [0, 1]$ is the mixing coefficient.

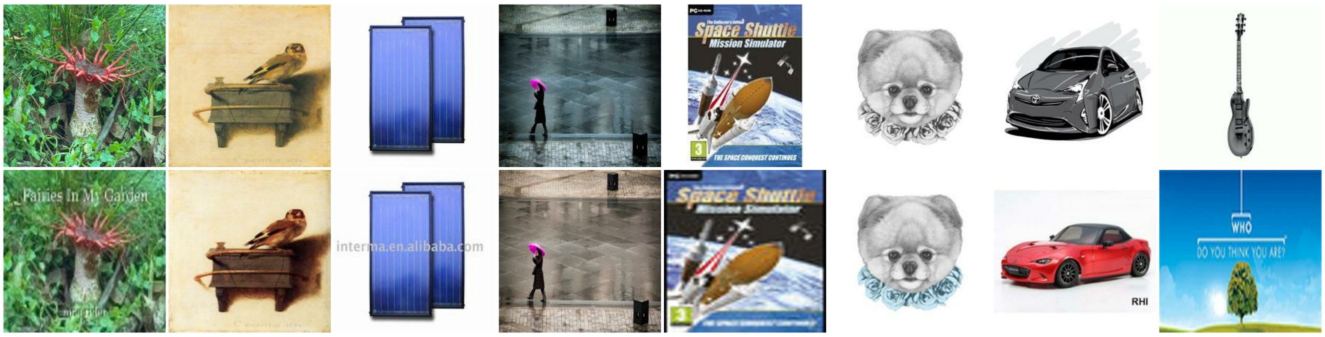
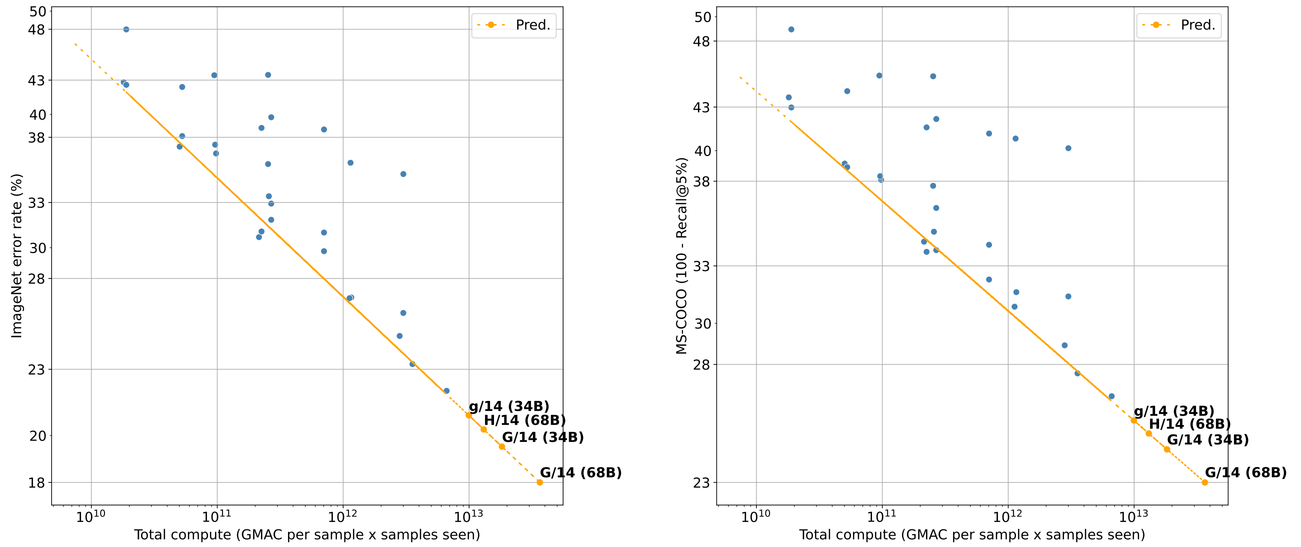


Figure 7. Duplicate images detected using pHash [82] between downstream datasets and LAION-400M. Top row shows images from downstream datasets, while bottom row show corresponding detected duplicates in LAION-400M. We observe near-duplicate detection for a variety of image transformations: blurring, text blitting, color transformations, cropping, and scaling. Last two columns show false positive examples detected on ImageNet-Sketch dataset. In general, we observed that most of false positive cases had a uniform background, which pHash seems to be sensitive to.



(a) Predictions on ImageNet

(b) Predictions on MS-COCO

Figure 8. Zero-shot performance extrapolation of g/14, H/14 and G/14 on larger scales. We fit a power-law on the Pareto frontier of available models. In Fig.8a we show the predictions for ImageNet classification, while in Fig.8b we show the predictions for MS-COCO image retrieval.

small, in the range of $0.2 - 0.5\%$ across different settings, which is small enough not to distort the trends observed in the effect of scale, where the changes are substantially larger.

LAION-400M and 400M subset of LAION-2B size. For 400M data scale, we are using LAION-400M dataset, as it was already validated by numerous previous works. This is not a subset of LAION-2B, as both were obtained by the same, but separately executed composition procedure using Common Crawl. To test that LAION-400M and LAION-2B can be considered as two different scale of the same data distribution, we extracted a random 400M subset from LAION-2B and conducted a pre-training experiment using our reference OpenCLIP ViT-B/32 model, 13B samples seen scale. We evaluated the pre-trained model on ImageNet zero-shot classification task, comparing it to same model pre-trained on LAION-400M. The outcome shows no significant difference between the performance of both models. This provides evidence that LAION-400M is comparable to a

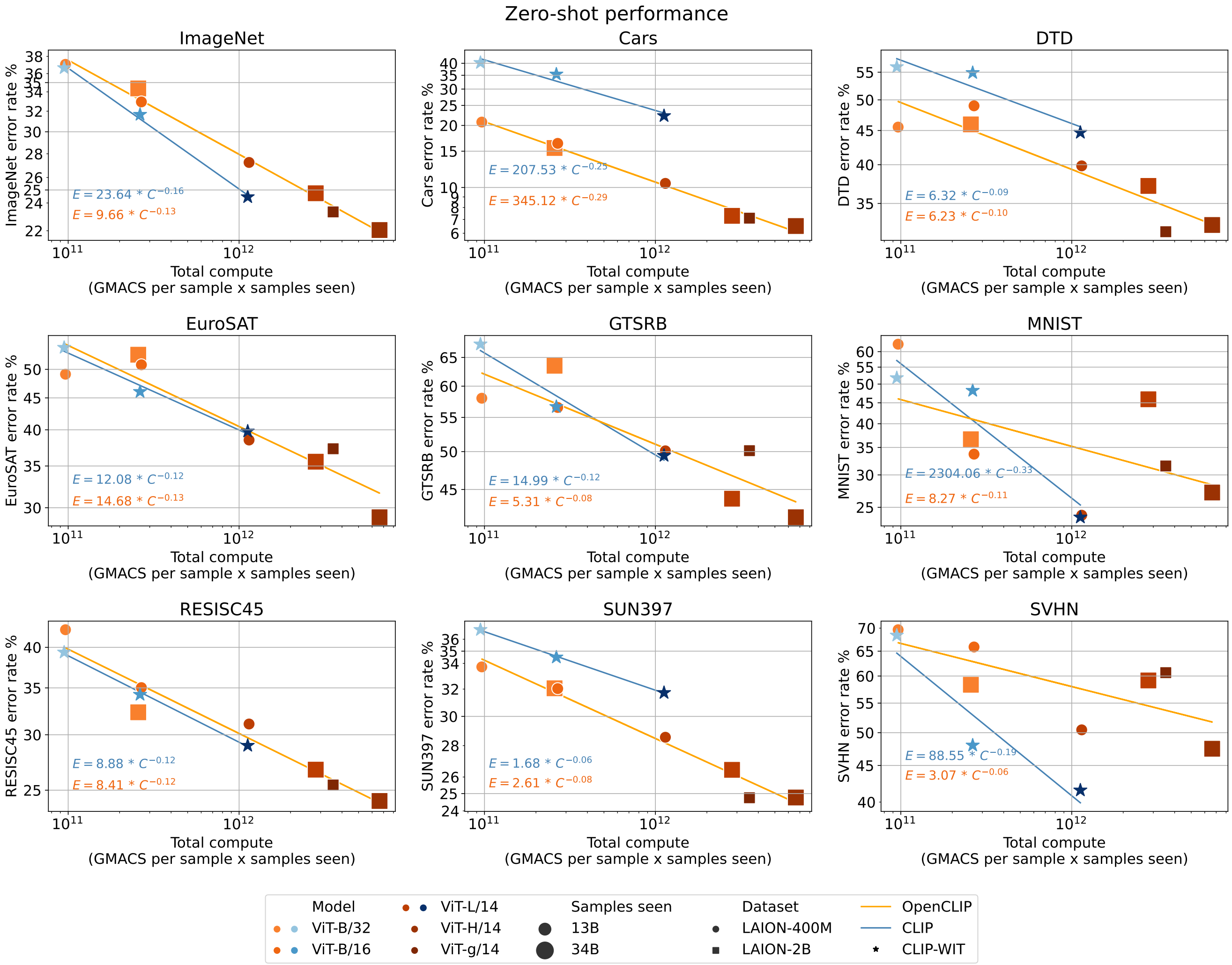


Figure 9. Scaling trends of zero-shot models on the eight other downstream tasks used for the fine-tuning experiments in Section 4.4 and on ImageNet.

Model	Samples Seen	Dataset	VTAB	ImageNet 10 shot	 25 shot	 Full	CIFAR100 10 shot	 25 shot	 Full
ViT-B/32	13B	CLIP-WIT	69.71	59.16	65.27	75.61	63.93	70.64	79.97
ViT-B/32	13B	LAION-400M	71.84	59.36	65.17	74.90	70.50	75.18	82.92
ViT-B/32	34B	LAION-2B	71.53	62.40	67.98	76.93	75.47	79.97	85.99
ViT-B/16	13B	CLIP-WIT	71.25	65.42	70.97	79.82	68.91	74.67	82.40
ViT-B/16	13B	LAION-400M	72.72	64.46	69.94	78.74	71.96	77.21	84.07
ViT-L/14	13B	CLIP-WIT	73.77	73.51	77.67	84.39	77.57	81.91	87.14
ViT-L/14	13B	LAION-400M	73.98	70.86	75.02	81.77	78.06	82.48	87.95
ViT-L/14	34B	LAION-2B	74.48	73.94	77.45	83.46	82.76	86.04	90.14
ViT-H/14	34B	LAION-2B	**75.96**	**75.79**	**79.07**	**84.85**	**84.74**	**87.82**	**91.43**
ViT-g/14	13B	LAION-2B	75.18	74.87	78.25	84.09	84.66	87.76	91.09

Table 5. Scaling model and data size leads to lower error linear classifers on ImageNet [15], CIFAR100 [43], and the visual task adaptation benchmark (VTAB) [85]. We train linear probes for models with at least 13B samples seen. We train probes by first caching the image features, thus no data augmentation is used. k shot denotes that k images per-class are used to train the linear probe.

400M subset extracted from LAION-2B, and can be thus considered to be a smaller scale of same data distribution.

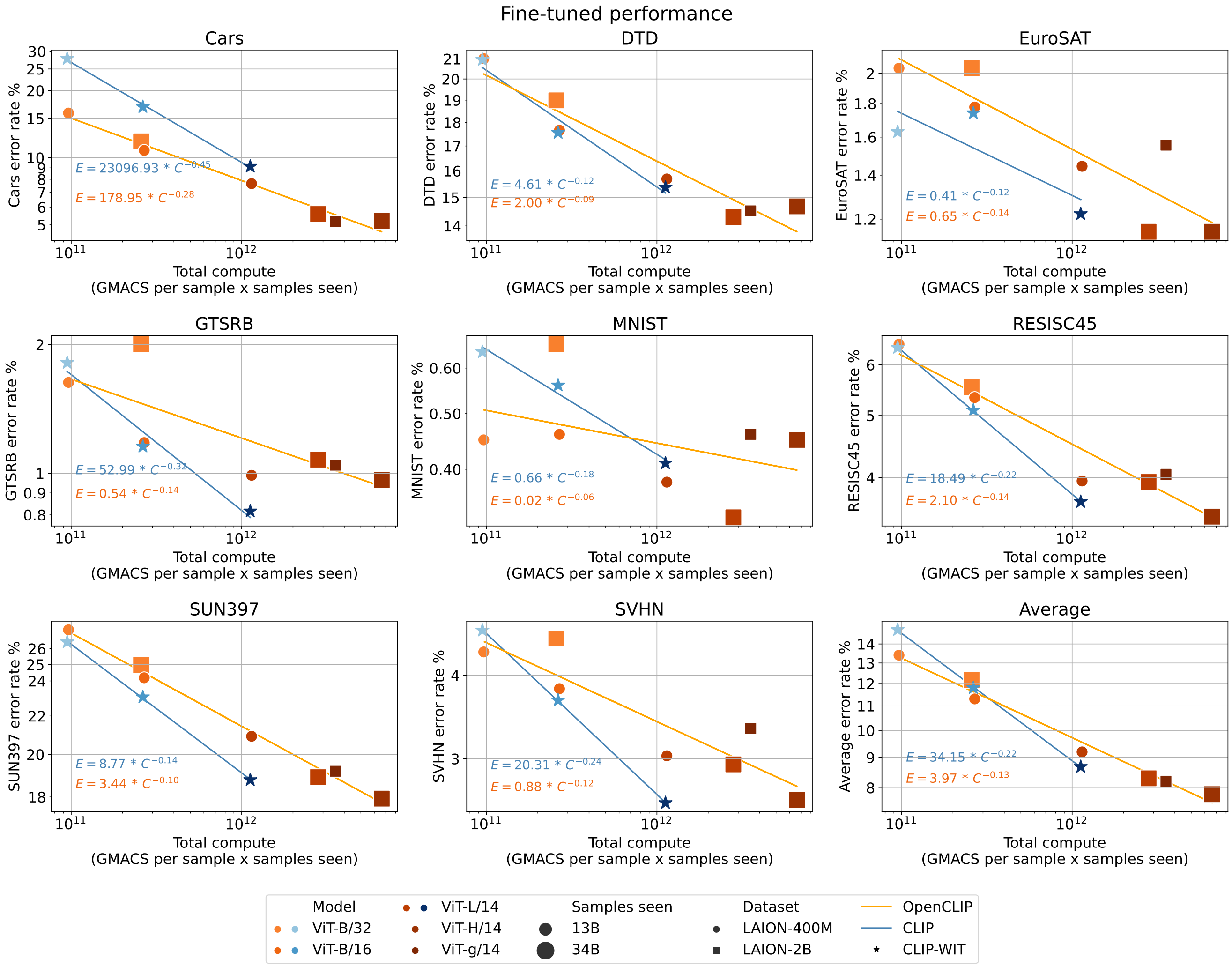


Figure 10. Scaling trends of fine-tuned models on the eight other downstream tasks used for the fine-tuning experiments in Section 4.4.

			Top-1 Accuracy (%)								
Arch.	# samples	Dataset	ImageNet	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC45	SUN397	SVHN
ViT-B/32	13B	CLIP-WIT	63.35	59.73	43.99	45.81	32.56	48.25	60.65	63.18	31.61
ViT-B/32	13B	LAION-400M	62.94	79.24	54.47	50.89	41.98	37.44	57.62	66.28	30.36
ViT-B/32	34B	LAION-2B	65.63	84.45	54.04	47.22	36.48	63.34	67.70	67.94	41.66
ViT-B/16	13B	CLIP-WIT	68.33	64.61	45.11	53.96	43.34	51.80	65.76	65.50	51.98
ViT-B/16	13B	LAION-400M	67.05	83.63	51.01	49.15	43.45	66.29	64.97	67.96	34.12
ViT-L/14	13B	CLIP-WIT	75.54	77.75	55.32	60.22	50.55	**76.36**	71.05	68.28	**58.45**
ViT-L/14	13B	LAION-400M	72.75	89.53	60.16	61.48	49.89	76.09	68.92	71.44	49.54
ViT-L/14	34B	LAION-2B	75.26	92.71	62.82	64.44	56.14	54.10	73.25	73.56	40.84
ViT-H/14	34B	LAION-2B	**77.95**	**93.50**	67.50	**71.04**	**58.35**	72.83	**75.87**	75.23	52.51
ViT-g/14	13B	LAION-2B	76.66	92.90	**68.24**	62.70	49.87	68.46	74.57	**75.24**	39.34

Table 6. Zero-shot accuracy for various models on downstream tasks from Section B.2.2.

Pre-training trial-to-trial variance. To have a sanity check of trial-to-trial variance for model pre-training, we trained our reference ViT-B/32 model, 13B samples seen scale, for two trials using exactly the same hyper-parameters (lr=0.001, batch size 86K, warm up 2K). We evaluated the two trials on ImageNet zero-shot classification task. The result suggests a small variance of around 0.1%, which is much smaller than variations observed when changing the scales. This allows us to conclude that scaling trends we observe are not distorted by variance caused by trial-to-trial

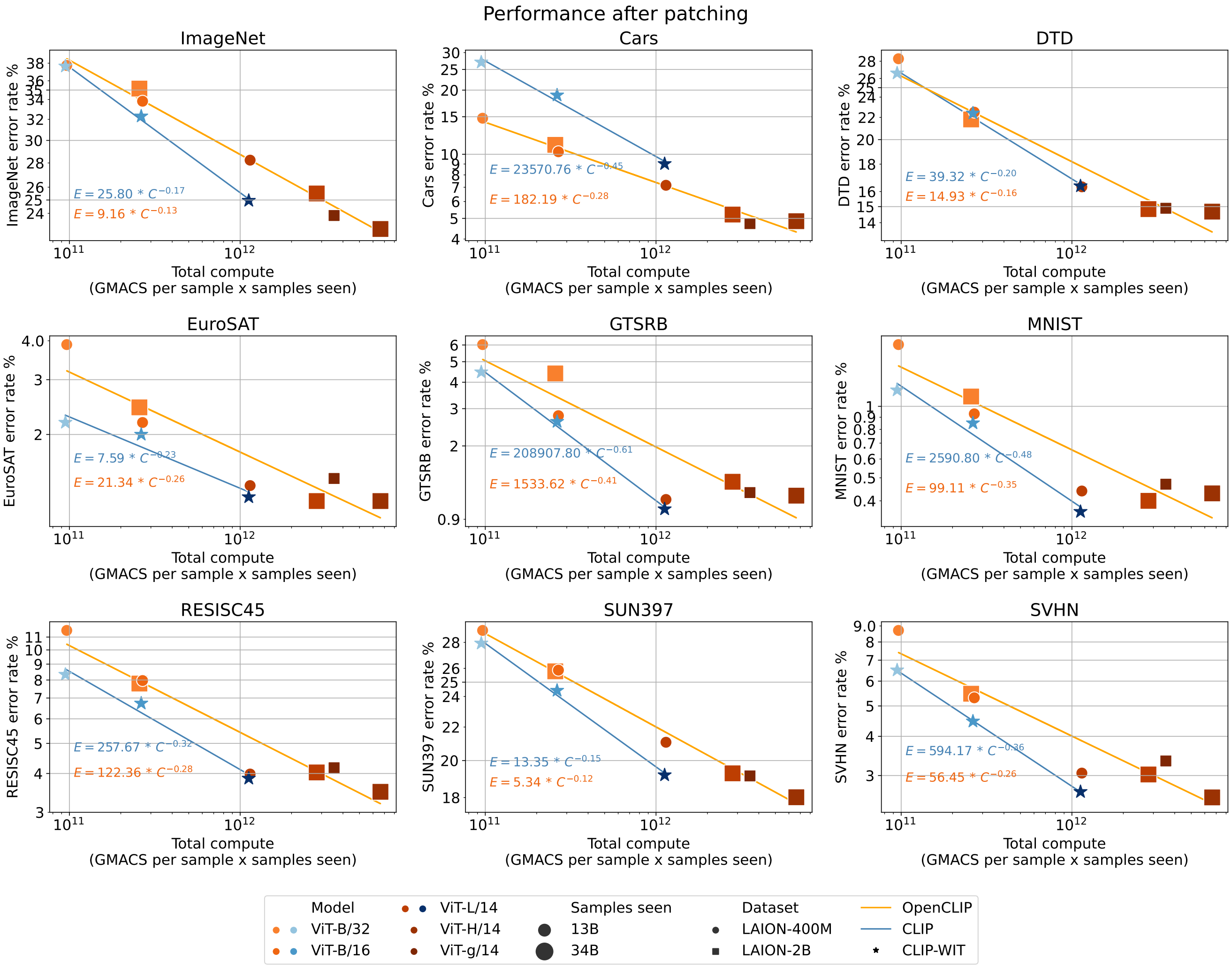


Figure 11. Scaling trends of patched models [31], on ImageNet and eight other downstream tasks used for the fine-tuning experiments in Section 4.4.

			Top-1 Accuracy (%)							
Arch.	# samples	Dataset	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC45	SUN397	SVHN
ViT-B/32	13B	CLIP-WIT	72.19	79.04	98.37	98.19	99.36	93.62	73.57	95.33
ViT-B/32	13B	LAION-400M	84.14	78.99	97.96	98.37	99.55	93.54	72.76	95.67
ViT-B/32	34B	LAION-2B	88.16	81.01	97.96	98.00	99.34	94.46	75.04	95.46
ViT-B/16	13B	CLIP-WIT	83.10	82.45	98.26	98.84	99.44	94.90	76.94	96.33
ViT-B/16	13B	LAION-400M	89.22	82.34	98.22	98.82	99.54	94.67	75.81	96.18
ViT-L/14	13B	CLIP-WIT	90.87	84.63	98.78	**99.18**	99.59	96.33	81.22	**97.42**
ViT-L/14	13B	LAION-400M	92.35	84.31	98.56	99.01	99.62	96.05	79.08	96.97
ViT-L/14	34B	LAION-2B	94.42	**85.69**	**98.85**	98.92	**99.67**	96.06	81.10	97.06
ViT-H/14	34B	LAION-2B	94.80	85.32	**98.85**	99.03	99.55	**96.52**	**82.08**	97.40
ViT-g/14	13B	LAION-2B	**94.84**	85.48	98.44	98.95	99.54	95.95	80.82	96.67

Table 7. Accuracy after fine-tuning for various models on downstream tasks from Section B.2.2. We fine-tune jointly on the eight downstream image classification tasks, alternating batches from each task. We fine-tune only the parameters of the vision encoder, using a fixed classification head for each task initialized with the weights from the zero-shot model.

pre-training.

Resampling vs full shuffled training. During our larger

			Top-1 Accuracy (%)								
Architecture	# samples	Dataset	ImageNet	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC45	SUN397	SVHN
ViT-B/32	13B	CLIP-WIT	62.36	72.96	73.40	97.81	95.53	98.83	91.67	72.08	93.50
ViT-B/32	13B	LAION-400M	62.27	85.24	71.70	96.11	93.97	98.18	88.44	71.03	91.29
ViT-B/32	34B	LAION-2B	64.84	88.93	78.19	97.56	95.60	98.90	92.21	74.22	94.53
ViT-B/16	13B	CLIP-WIT	67.70	81.07	77.61	98.00	97.40	99.15	93.27	75.59	95.53
ViT-B/16	13B	LAION-400M	66.18	89.73	77.50	97.81	97.22	99.07	92.03	74.13	94.69
ViT-L/14	13B	CLIP-WIT	75.04	90.98	83.62	98.74	**98.99**	**99.64**	96.14	80.81	97.34
ViT-L/14	13B	LAION-400M	71.76	92.85	83.67	98.63	98.88	99.56	96.02	78.93	96.95
ViT-L/14	34B	LAION-2B	74.50	94.79	85.16	**98.78**	98.65	99.60	95.97	80.72	96.98
ViT-H/14	34B	LAION-2B	**77.12**	95.15	**85.32**	**98.78**	98.84	99.57	**96.51**	**81.98**	**97.45**
ViT-g/14	13B	LAION-2B	76.16	**95.27**	85.11	98.56	98.80	99.53	95.83	80.86	96.66

Table 8. Accuracy after joint patching [31] for various models on downstream tasks from Section B.2.2. Patching by jointly fine-tuning on the eight tasks with the exception of ImageNet (used only as control), then interpolating the weights of the fine-tuned model with the weights of the zero-shot model. The mixing coefficient for the interpolation is chosen so it maximizes average accuracy on the eight downstream tasks while maintaining ImageNet accuracy within 1 percentage point of the corresponding zero-shot model.

							Top-1 Accuracy (%)					
Model	Im Size	Dataset	Extra FT	Params (M)	GMAC	Acts (M)	IN	IN-ReaL	IN-V2	IN-A	IN-R	IN-Sketch
ViT-B/32	224	CLIP-WIT	None	88.2	4.4	5.0	81.93	87.17	70.70	22.57	55.90	45.04
ViT-B/32	224	LAION-2B	None	88.2	4.4	5.0	82.58	87.54	71.21	22.85	59.16	49.07
ViT-B/32	224	LAION-2B	IN-12k	88.2	4.4	5.0	83.30	87.81	72.50	30.57	57.06	45.74
ViT-B/32	384	CLIP-WIT	IN-12k	88.3	13.1	16.5	85.11	89.04	74.53	44.75	58.21	45.75
ViT-B/16	224	CLIP-WIT	None	86.6	17.6	23.9	85.28	89.16	75.57	47.23	66.02	50.94
ViT-B/32	384	LAION-2B	IN-12k	88.3	13.1	16.5	85.38	89.20	75.08	47.95	60.37	47.95
ViT-B/16	224	LAION-2B	None	86.6	17.6	23.9	85.47	89.43	75.13	41.57	68.75	55.40
ViT-B/16	384	CLIP-WIT	None	86.9	55.5	101.6	86.24	89.71	76.68	57.55	67.22	52.15
ViT-B/16	384	LAION-2B	None	86.9	55.5	101.6	86.53	90.04	77.55	56.96	69.94	55.85
ViT-B/16	384	LAION-2B	IN-12k	86.9	55.5	101.6	87.17	90.11	78.16	62.61	65.53	52.62
ViT-L/14	224	LAION-2B	None	304.2	81.1	88.8	87.30	90.10	78.42	59.89	81.70	64.81
ViT-H/14	224	LAION-2B	None	632.0	167.4	139.4	87.59	90.17	79.36	65.56	**83.28**	**67.41**
ViT-L/14	336	LAION-2B	None	304.5	191.1	270.2	87.78	90.30	79.07	69.03	82.60	64.79
ViT-L/14	224	CLIP-WIT	None	304.2	81.1	88.8	87.85	90.31	**79.59**	71.79	82.32	62.63
ViT-L/14	224	LAION-2B	IN-12k	304.2	81.1	88.8	87.89	90.30	78.51	67.01	78.26	62.06
ViT-L/14	336	LAION-2B	IN-12k	304.5	191.1	270.2	88.17	90.43	78.84	73.64	77.68	60.97
ViT-L/14	224	CLIP-WIT	IN-12k	304.2	81.1	88.8	88.17	90.37	79.38	72.33	78.68	61.40
ViT-H/14	224	LAION-2B	IN-12k	632.0	167.4	139.4	88.25	90.41	79.22	70.72	82.82	65.32
ViT-H/14	336	LAION-2B	IN-12k	632.5	391.0	407.5	**88.50**	**90.49**	79.55	**75.68**	82.26	64.62

Table 9. Fine-tune results for ImageNet-1k and associated robustness test sets (ImageNet-ReaL [7], ImageNet-V2 [60], ImageNet-A [24], Imagenet-R [22], and ImageNet-Sketch [75]). Rows with the 'Extra FT' set to IN-12k were fine-tuned on a 12k class subset of ImageNet-22k before fine-tuning on ImageNet.

scale pre-training experiments featuring LAION-2B, it became important to allow for frequent checkpoint saving. Saving within a running epoch would require to memorize which samples were already seen, to be able to resume training in such a way that only previously not seen samples would be taken. To simplify the procedure, we have tested a version that does not perform epoch-wise training, taking a pre-defined number of samples instead for a virtual ”step” through data. Such a resampling procedure can have repeated samples in the subset of data that contains in total the number of samples equal to number of samples in one full epoch through the dataset. As such training procedure differs from standard epoch-wise training, we conducted test experiments to check whether this results in differences in performance of pre-trained models when comparing to standard epoch-wise shuffling training. We trained our reference ViT-B/32 model and ViT-B/16 model on LAION-400M either using standard epoch-wise training with shuffling or the training that involves described resampling procedure. We observed only negligible differences of 0.1%-0.3%, concluding that using simple resampling cannot distort scaling trends observed in the study.

B.2.4 Further detailed results and analysis

Consistency of scaling trends for CLIP and openCLIP

In Fig. 12, we complement the results we found in Fig. 1 and show scaling trends with additional ResNet models from OpenAI trained on the WebImageText (WIT) dataset. De-

Hyperparameter	B/32	B/16	L/14	H/14
Peak Learning-rate	1.00E-03	3.00E-04	6.00E-05	5.00E-05
Batch Size	4096	2048	2048	2048
Epochs	50	50	50	50
Warmup Epochs	10	10	10	10
Layer-wise LR Decay	0.65	0.7	0.8	0.82
EMA Weight Smoothing	0.9998	0.9998	0.9997	0.9998
Weight Decay	0.05	0.05	0.01	0.02
Label Smoothing	0.1	0.1	0.1	0.1
Stoch. Depth	0.1	0.1	0.2	0.2
Dropout	0	0	0	0
Gradient Clipping	3	3	3	2
Rand Augment (Uniform)	M=U(0, 8), N=2	M=U(0, 9), N=2	M=U(0, 9), N=3	M=U(0, 8), N=4
Random Erase Prob	0.3	0.3	0.3	0.3
Random Resize Crop	Yes	Yes	Yes	Yes
Mixup Alpha	0	0	0	0
Cutmix Alpha	0	0	0	0
Color Jitter	0	0	0	0

Table 10. ImageNet fine-tune hyper-parameters.

Hyperparameter	B/32	B/16	L/14	H/14
Peak Learning-rate	1.00E-03	5.00E-04	5.00E-04	4.00E-04
Batch Size	4096	4096	4096	4096
Epochs	60	60	60	60
Warmup Epochs	10	10	10	10
Layer-wise LR Decay	0.65	0.7	0.8	0.86
EMA Weight Smoothing	0.9998	0.9998	0.9999	0.9999
Weight Decay	0.05	0.05	0.02	0.02
Label Smoothing	0.1	0.1	0.1	0.1
Stoch. Depth	0.1	0.1	0.2	0.2
Dropout	0	0	0	0
Gradient Clipping	3	3	3	2
Rand Augment (Uniform)	M=U(0, 8), N=2	M=U(0, 8), N=2	M=U(0, 9), N=2	M=U(0, 8), N=3
Random Erase Prob	0.3	0.3	0.3	0.3
Random Resize Crop	Yes	Yes	Yes	Yes
Mixup Alpha	0	0	0	0
Cutmix Alpha	0	0	0	0
Color Jitter	0	0	0	0

Table 11. ImageNet-12k intermediate fine-tune hyper-parameters.

Batch size / Model	32k/38k (L/14)	64k (B/16) / 86k (+lr tune)
ViT B/32	62.9	63.37
ViT B/16	67.34	67.86
ViT L/14	72.8	72.98

Table 12. Batch size control experiments, zero-shot ImageNet top-1 accuracy. Executed on LAION-400M, 13B samples seen (32 full epochs).

spite different model architectures, we see the same consistent pattern - both OpenAI's ResNets and ViT trained on WIT demonstrate stronger scaling on zero-shot classification but worse scaling on retrieval than openCLIP trained on LAION datasets. This evidence further backs up our findings on dataset-dependent scaling law differences across tasks we observe for LAION-openCLIP and WIT-openAI CLIP

Model/Dataset	400M LAION-2B subset	LAION-400M
ViT B/32	63.56	63.37

Table 13. 400M data scale subset control experiments, zero-shot ImageNet top-1 accuracy. Executed either on 400M subset of LAION-2B or on LAION-400M, 13B samples seen (32 full epochs).

Trial	ImageNet zero-shot top-1
1	63.28
2	63.67

Table 14. Trial-to-trial variance control experiment. Executed on LAION-400M, 13B samples seen (32 full epochs) using ViT B/32 model.

Model	Resampling	Full shuffling
ViT B/32	63.37	63.28; 63.67

Table 15. Resampling vs. full shuffling control experiments, zero-shot ImageNet top-1 accuracy. Executed on LAION-400M, 13B samples seen (32 full epochs).

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	38.05	41.53	43.66
	13B	42.30	46.18	45.50
	34B	42.10	46.41	50.69
ViT-B/16	3B	43.48	45.14	46.93
	13B	44.42	48.39	48.72
	34B	44.45	48.31	52.60
ViT-L/14	3B	45.69	50.50	51.64
	13B	46.36	51.51	53.01
	34B	45.70	52.83	54.63
ViT-H/14	34B	-	-	56.43
ViT-g/14	13B	-	-	56.54

Table 16. Detailed results on VTAB+ [65] zero-shot classification, where we average over 35 tasks.

models (see also Discussion in Sec. 5).

Details of zero-shot classification results. Complementing results from the Section 4, we provide summary tables for the performance measure on different downstream tasks: ImageNet (Tab. 19), ImageNet robustness(Tab. 20), MS-COCO image retrieval (Tab. 21) and text retrieval (Tab. 22), Flickr30K image retrieval (Tab. 23) and text retrieval (Tab. 24), and VTAB+ (Tab. 16 and 17).

Details of linear probing results. To supplement Figures 2 and 3, we provide the corresponding Table 5 with detailed results.

Architecture and training hyperparameters. We provide overview for architecture (Tab. 25) and pre-training hyper-parameters (Tab. 18) that we have used in the experiments.

C. Code and Data availability

We provide source code used for running experiments and producing figures in this study at https://github.com/LAION-AI/scaling-laws-openclip. Links to pre-trained models obtained in this study and links to instructions for obtaining LAION-400m and LAION-5B used for pre-training experiments will be also made available there. All datasets used in the study are openly available and are listed together with references to the original work in Table 26.

Broader and Social Impact

Safety aspect. Our work deals with studying function and properties of pre-trained models on large scales. Releasing these models to public can have both positive and negative implications, like with any research artefact that possesses generic functionality. We would like to stress that we consider the released pre-trained language-vision models as research artefacts that are there to advance the studies of scaling laws and allow analysis of the properties and behavior of such models for the broader research community. These models are not meant to be incorporated into end products or even used for applications in sensitive areas like interpretation of medical imaging in hospitals or security surveillance. There is potential for abuse of technology based on large-scale pre-trained generalist models, and it is the task of democratic institutions to work out rules for sensitive applications that might involve those. Open release of models gives the broad research community also opportunity to study safety related aspects of such models, such to preventively design measures that make such abuse by malicious parties less probable, in a common transparent effort. Same applies to the common effort of studying yet not systematically understood biases that such models may contain due to pre-training on either largely uncurated, imbalanced data or on data filtered by models that already contain unknown biases (like OpenAI's CLIP that was trained on the private WIT-400M dataset), and due to the simplistic nature of the contrastive InfoNCE loss that drives learning.

Energy cost. There is high computational cost bound to pre-training experiments on large scale. Supercomputers used in our studies are highly ranked in the Green Top-500 list, ensuring that energy costs are dampened. In addition, strongly transferable pre-trained models save energy on numerous downstream tasks where they can perform in data-efficient and thus in an energy saving manner. Releasing such pre-trained models to public incurs additional energy savings, as research community can re-use already validated models without necessity to train those from scratch again.

Dataset	B/32 (34B)	B/16 (34B)	L/14 (34B)	g/14 (13B)	H/14 (34B)
INet	66.47	70.22	75.20	76.66	**77.97**
INet-v2	58.16	62.28	67.69	69.61	**70.82**
INet-R	76.47	80.59	87.41	88.65	**89.32**
INet-S	53.72	56.09	63.28	65.22	**66.57**
ObjNet	48.78	56.05	65.50	67.47	**69.70**
INet-A	25.43	38.23	53.88	57.11	**59.23**
CIFAR-10	93.65	94.94	96.64	97.05	**97.42**
CIFAR-100	75.47	76.83	83.36	83.91	**84.68**
MNIST	67.73	65.99	54.87	69.04	**72.94**
Flowers102	72.35	71.23	75.90	77.61	**80.21**
Cars	86.15	88.50	92.61	92.77	**93.46**
SVHN	43.51	51.39	46.30	**60.33**	56.13
FER2013	46.02	51.78	**53.71**	46.57	51.76
RenderedSST2	57.17	59.80	59.31	**64.58**	64.09
Pets	89.81	90.52	93.21	94.28	**94.39**
Caltech-101	83.50	83.83	85.04	**85.22**	85.04
VOC2007-Cl	79.75	78.85	80.52	**81.03**	77.61
SUN397	68.57	70.85	74.33	**75.40**	75.22
FGVC Aircraft	24.06	27.00	36.93	37.80	**42.75**
Country211	16.78	20.31	26.36	28.73	**30.01**
DTD	55.64	56.33	62.77	**68.14**	67.87
GTSRB	49.49	48.24	56.10	49.74	**58.45**
STL10	96.55	97.86	**98.86**	98.59	98.44
Retino	**73.42**	67.96	21.06	43.42	23.80
EuroSAT	46.94	53.46	65.15	64.80	**71.74**
RESISC45	60.71	62.76	66.67	**71.71**	69.57
PCAM	**59.44**	56.37	55.26	55.09	53.63
CLEVR Counts	15.02	21.49	31.09	**33.19**	27.84
CLEVR Dist	14.54	**21.07**	16.10	17.73	16.77
DSPRITES Orient	**3.77**	2.68	2.00	3.08	2.61
DSPRITES pos	2.80	3.30	3.15	**3.54**	3.14
SmallNORB Elv	**11.70**	11.30	10.95	11.34	11.13
SmallNORB Azim	5.86	5.67	5.63	**5.88**	5.50
DMLAB	17.48	19.93	**22.43**	19.02	14.20
KITTI Dist	**27.14**	17.16	22.93	14.63	11.11
VTAB+ (Avg.)	50.69	52.60	54.63	**56.54**	56.43

Table 17. Detailed zero-shot top-1 classification results of LAION-2B models on VTAB+ 35 tasks. We highlight the best results for each downstream dataset.

Model	Dataset	BS. (global)	LR.	Warm.	#samples.	#GPUs	Time (hrs.)	GPU-h/MWh
B/32	LAION-80M	256(32768)	5e-4	2K	3B	128	7	836/0.29
B/32	LAION-80M	256(32768)	5e-4	2K	13B	128	33	4181/1.46
B/32	LAION-80M	256(88064)	1e-3	10K	34B	344	96	32953/11.53
B/32	LAION-400M	256(88064)	1e-3	10K	3B	344	3	1063/0.37
B/32	LAION-400M	672(86016)	1e-3	2K	13B	128	70	8912/3.12
B/32	LAION-400M	256(32768)	5e-4	2K	34B	128	87	11177/3.91
B/32	LAION-2B	256(88064)	1e-3	10K	3B	344	3	1121/0.39
B/32	LAION-2B	256(32768)	5e-4	2K	13B	128	39	4954/1.73
B/32	LAION-2B	96(79104)	1e-3	2K	34B	824	51	42307/14.81
B/16	LAION-80M	256(88064)	1e-3	10K	3B	344	6	1900/0.66
B/16	LAION-80M	512(90112)	1e-3	10K	13B	176	71	12518/4.38
B/16	LAION-80M	256(88064)	1e-3	10K	34B	344	70	24032/8.41
B/16	LAION-400M	256(88064)	1e-3	10K	3B	344	5	1713/0.60
B/16	LAION-400M	192(33792)	5e-4	10K	13B	176	61	10736/3.76
B/16	LAION-400M	512(90112)	1e-3	10K	34B	176	148	26009/9.10
B/16	LAION-2B	256(88064)	1e-3	10K	3B	344	5	1822/0.64
B/16	LAION-2B	512(90112)	1e-3	10K	13B	176	66	11675/4.09
B/16	LAION-2B	256(88064)	1e-3	10K	34B	344	121	41726/14.60
L/14	LAION-80M	224(88704)	1e-3	10K	3B	396	18	7243/2.54
L/14	LAION-80M	448(89600)	1e-3	10K	13B	200	102	20393/7.14
L/14	LAION-80M	224(89600)	1e-3	10K	34B	400	227	90647/31.73
L/14	LAION-400M	224(88704)	1e-3	10K	3B	396	17	6717/2.35
L/14	LAION-400M	112(86016)	1e-3	2K	13B	768	61	46735/16.36
L/14	LAION-400M	84(86016)	1e-3	10K	34B	1024	122	124727/43.65
L/14	LAION-2B	224(88704)	1e-3	10K	3B	396	18	7055/2.47
L/14	LAION-2B	84(86016)	1e-3	10K	13B	1024	52	53599/18.76
L/14	LAION-2B	224(86016)	1e-3	10K	34B	384	319	122509/42.88
H/14	LAION-2B	96(79104)	5e-4	2K	34B	824	279	229665/80.38
g/14	LAION-2B	80(64000)	5e-4	2K	13B	800	137	109392/38.29
Total								1058318/370.41

Table 18. Training hyper-parameters and resources used to for pre-training our models on LAION 80M, 400M, and 2B subsets. Note that **BS** refer to batch size per GPU worker (with **global** the corresponding global batch size), **LR** to base learning rate, **Warm** to the total number of warmup steps, **Time** to total training time in hours, **GPU-h** to GPU hours, **MWh** to the total energy consumed in Megawatt hours.

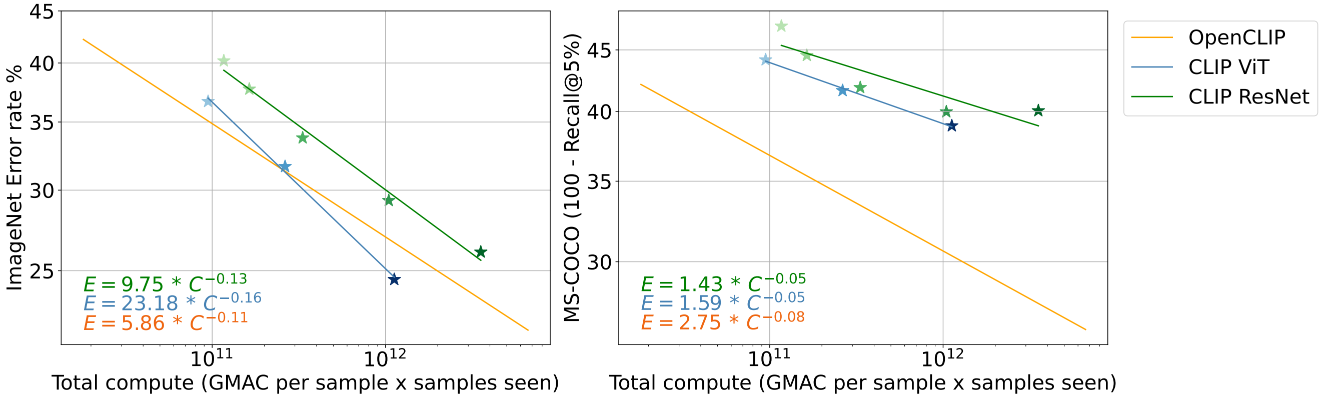


Figure 12. Consistency of scaling trends for CLIP and openCLIP with additional CLIP ResNet models from OpenAI and trained on WIT. Relationship between total compute (GMAC) and ImageNet zero-shot classification error rate **(Left)** and with MS-COCO image retrieval at Recall@5 **(Right)**.

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	51.94	57.12	57.36
	13B	56.46	63.23	62.53
	34B	56.43	64.06	66.47
ViT-B/16	3B	57.55	62.68	61.82
	13B	60.24	67.00	68.13
	34B	61.28	69.00	70.22
ViT-L/14	3B	61.14	69.31	68.93
	13B	63.96	73.06	73.10
	34B	64.83	73.94	75.20
ViT-H/14	34B	-	-	77.97
ViT-g/14	13B	-	-	76.66

Table 19. Detailed results on ImageNet zero-shot accuracy.

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	37.95	41.60	42.44
	13B	42.23	48.97	48.83
	34B	43.01	50.12	52.51
ViT-B/16	3B	43.48	47.82	48.07
	13B	47.29	54.89	55.89
	34B	49.29	57.14	58.65
ViT-L/14	3B	48.26	57.53	57.56
	13B	52.23	63.84	64.61
	34B	54.23	65.25	67.55
ViT-H/14	34B	-	-	71.13
ViT-g/14	13B	-	-	69.61

Table 20. Detailed results on ImageNet five robustness datasets zero-shot accuracy (average over the five datasets is reported).

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	51.04	56.29	57.01
	13B	54.67	61.90	61.66
	34B	54.72	62.28	65.05
ViT-B/16	3B	55.83	60.85	61.08
	13B	57.83	63.64	66.11
	34B	58.84	65.81	67.73
ViT-L/14	3B	58.42	65.63	66.21
	13B	59.18	68.40	69.16
	34B	59.84	68.62	71.08
ViT-H/14	34B	-	-	73.43
ViT-g/14	13B	-	-	72.40

Table 21. Detailed results on MS-COCO image retrieval Recall@5.

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	67.16	73.38	73.10
	13B	70.32	77.60	77.04
	34B	70.78	77.46	79.58
ViT-B/16	3B	72.22	77.18	76.72
	13B	73.84	79.62	81.00
	34B	74.12	80.52	81.78
ViT-L/14	3B	74.90	80.78	79.86
	13B	76.24	82.12	82.94
	34B	75.96	83.44	84.00
ViT-H/14	34B	-	-	86.04
ViT-g/14	13B	-	-	85.36

Table 22. Detailed results on MS-COCO text retrieval Recall@5.

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	76.00	80.50	82.16
	13B	78.46	85.20	85.36
	34B	78.98	85.90	88.26
ViT-B/16	3B	80.78	85.84	85.12
	13B	84.76	88.16	89.90
	34B	84.38	89.58	90.32
ViT-L/14	3B	84.16	89.14	89.82
	13B	84.86	91.04	91.72
	34B	85.70	91.28	92.92
ViT-H/14	34B	-	-	94.10
ViT-g/14	13B	-	-	93.48

Table 23. Detailed results on Flickr30K image retrieval Recall@5.

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	88.20	91.60	92.70
	13B	91.30	95.60	94.50
	34B	90.70	95.60	96.10
ViT-B/16	3B	91.90	95.60	94.60
	13B	94.90	96.80	97.60
	34B	94.80	97.40	98.00
ViT-L/14	3B	93.60	97.80	96.70
	13B	95.00	98.30	98.40
	34B	96.90	97.70	98.70
ViT-H/14	34B	-	-	99.30
ViT-g/14	13B	-	-	99.10

Table 24. Detailed results on Flickr30K text retrieval Recall@5.

Name	Width	Emb.	Depth	Acts.	Params	GMAC
ViT-B/32	768 / 512	512	12 / 12	10 M	151 M	7.40
ViT-B/16	768 / 512	512	12 / 12	29 M	150 M	20.57
ViT-L/14	1024 / 768	768	24 / 12	97 M	428 M	87.73
ViT-H/14	1280 / 1024	1024	32 / 24	161 M	986 M	190.97
ViT-g/14	1408 / 1024	1024	40 / 24	214 M	1.37 B	290.74
ViT-G/14	1664 / 1280	1280	48 / 32	310 M	2.54 B	532.92

Table 25. Hyper-parameters of different architectures we consider. **Emb** refers to embedding size, **Acts** refers to the number of activations in millions, and **Params** refers to the number of parameters in millions. **GMAC** refers to giga multiply–accumulates. All entries in the form of A / B denote image and text parameters respectively.

Dataset	Abbr.	Test size	#Classes
ImageNet	INet	50,000	1,000
ImageNet-v2	INet-v2	10,000	1,000
ImageNet-R	INet-R	30,000	200
ImageNet Sketch	INet-S	50,889	1,000
ObjectNet	ObjNet	18,574	113
ImageNet-A	INet-A	7,500	200
CIFAR-10	-	10,000	10
CIFAR-100	-	10,000	100
MNIST	-	10,000	10
Oxford Flowers 102	Flowers102	6,149	102
Stanford Cars	Cars	8,041	196
SVHN	-	26,032	10
Facial Emotion Recognition 2013	FER2013	7,178	7
RenderedSST2	-	1,821	2
Oxford-IIIT Pets	Pets	3,669	37
Caltech-101	-	6,085	102
Pascal VOC 2007 Classification	VOC2007-Cl	14,976	20
SUN397	-	108,754	397
FGVC Aircraft	-	3,333	100
Country211	-	21,100	211
Describable Textures	DTD	1,880	47
GTSRB	-	12,630	43
STL10	-	8,000	10
Diabetic Retinopathy	Retino	42,670	5
EuroSAT	-	5,400	10
RESISC45	-	6,300	45
PatchCamelyon	PCAM	32,768	2
CLEVR Counts	-	15,000	8
CLEVR Object Distance	CLEVR Dist	15,000	6
DSPRITES Orientation	DSPRITES Orient	73,728	40
DSPRITES Position	DSPRITES pos	73,728	32
SmallNORB Elevation	SmallNORB Elv	12,150	9
SmallNORB Azimuth	SmallNORB Azim	12,150	18
DMLAB	-	22,735	6
KITTI closest vehicle distance	KITTI Dist	711	4
MS-COCO	-	5,000	-
Flickr30K	-	1,000	-

Table 26. Datasets used for evaluating downstream performance. Adapted from [65].

Acknowledgments.

We would like to express gratitude to all the people who are working on making code, models and data publicly available, advancing community based research and making research more reproducible. Specifically, we would like to thank all the members of the LAION discord server[5] community that was pivotal for the effort to compose LAION-400m and LAION-5B datasets without which this study would be impossible, and openAI for making their pre-trained CLIP models publicly available. We want to thank Hugging Face for providing hosting space for open datasets and models and Stability AI for providing supercomputing resources and storage space.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. [6] for funding this work by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS Booster [34] at Jülich Supercomputing Centre (JSC). We also acknowledge storage resources on JUST [20] granted and operated by JSC, as well as computing resources from the Helmholtz Data Federation (HDF). Further thank goes for support provided by JSC supercomputing facility administration team, especially to Damian Alvarez for his endurance and patience during the long ”de-micing” sessions on JUWELS Booster.

Special thanks goes also to Richard Vencu (LAION, Stability AI) for his on-going dedication towards enabling a HPC system and infrastructure around it that can be used by broad community of researchers and citizen scientists.

Author Contributions

- **Mehdi Cherti**: Planned and executed experiments on JUWELS Booster and Stability AI HPC, coordinated and performed results collection, distillation and analysis, performed deduplication analysis on LAION and downstream targest datasets, manuscript writing and revision.

- **Romain Beaumont**: Planned and executed experiments on JUWELS Booster and Stability AI HPC, pioneered training of large scale openCLIP ViT H-14 and g-14, manuscript revision.

- **Ross Wightman**: Planned and executed experiments on JUWELS Booster and Stability AI HPC, performed full fine-tuning experiments, manuscript writing and revision.

- **Mitchell Wortsman**: experiment design and planning, performed experiments evaluating linear probing finetuning performance and robustness on ImageNet and other downstream datasets, manuscript writing and revision.

- **Gabriel Ilharco**: experiment design and planning, performed experiments evaluating full fine-tuning performance and robustness on ImageNet and other downstream datasets, manuscript writing and revision.

- **Cade Gordon**: implemented local contrastive loss for efficient distributed training, manuscript writing and revision.

- **Christoph Schuhmann**: supervised larger scale experiments, resource and community organization, compute and storage resource acquisition, manuscript revision.

- **Ludwig Schmidt**: provided advice on experiment design and study directions, manuscript writing and revision, general supervision.

- **Jenia Jitsev**: led the project; conducted experiments on various scales using JUWELS Booster and Stability AI HPC, scientific organization & manuscript writing, ethical and social content, experiments planning and design, compute and storage resource acquisition, general supervision.

[5] https://discord.gg/BZqhreFazY

[6] https://gauss-centre.eu