

Appendix: Scaling Laws for Comparison of Open Foundation Language-Vision Models and Datasets

A Estimated parameters for scaling law fits

To complement main results for scaling law based comparison of CLIP and MaMMUT (Sec. 3.1, Fig. 1, 2), we provide exact numbers of scaling law fits for both openCLIP and openMaMMUT measurements on zero shot IN1K classification and MS-COCO retrieval downstream tasks (Tab. 5a). Estimated values of exponents in power laws alone do not tell which models are more scalable, as we use here the functional form with additive terms for both irreducible error and non-zero random model performance (Eq. 1). Apart from plot visualization attesting Mammut stronger scalability than CLIP (Fig. 1, 2), scalability can be also compared via computing derivatives of the obtained fit in selected compute points. Derivatives that have larger absolute values stand for larger slope (bigger rate of decrease) and thus indicate stronger scalability. In Tab. 5b we show derivatives computed for scaling law fits, obtaining larger derivatives for openMaMMUT than for openCLIP, confirming again stronger scalability for MaMMUT over CLIP.

B More details on scaling law derivation experiments

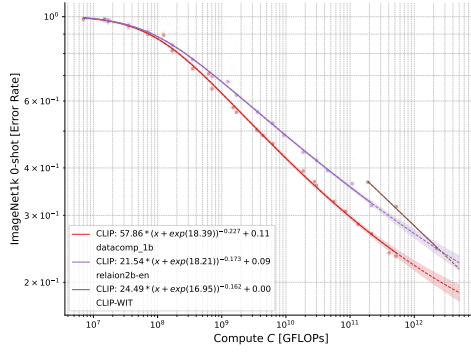
Compute budget and energy consumption for the experiments. In Tab. 4, we provide overview over the GPU hours and energy spent for scaling law derivation experiments. We provide separate calculation for different learning rate schedule types (cosine, constant learning rate and constant learning rate + cooldown), for different datasets (Re-LAION-1.4B and DataComp-1.4B) and for different GPU types (A100 and H100). Large fraction of resources was spent for reference cosine schedule based scaling law derivation on DataComp-1.4B. We see that despite higher density of possible measurements, const based schedules use substantially less compute.

Detailed versions of scaling law plots. In the more detailed versions of scaling law plots (Fig. 16 and 17) we see the separate scaling curves for each model size (cooler colors indicate smaller models). The **bigger models require larger sample seen scale** to unfold their performance advantage, with the performance lagging behind smaller scale models on same smaller compute scale, where larger models suffer from sample seen scale bottleneck. On the other hand, **for the higher compute and samples seen scales, smaller models tend to saturate**, indicating a bottleneck in model number of parameters.

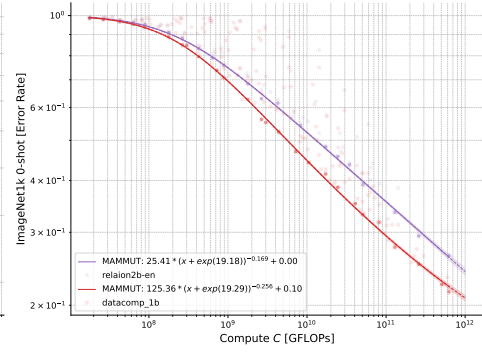
Data efficiency on Re-LAION and DFN. As we see from Fig. 11, MaMMUT exhibits consistently more superior scaling with respect on training data size on both Re-LAION-1.4B and DFN-1.4B. This supports the conclusion that **MaMMUT is more data efficient across multiple training datasets**.

LR Scheduler	GPU	Dataset	MWh	GPU Hours
NVIDIA A100				
cosine	NVIDIA-A100	DataComp-1.4B	2.59e+05	1.03e+06
const-cooldown	NVIDIA-A100	DataComp-1.4B	1.43e+05	5.72e+05
const	NVIDIA-A100	DataComp-1.4B	9.30e+04	3.72e+05
cosine	NVIDIA-A100	Re-LAION-1.4B	3.91e+04	1.56e+05
const-cooldown	NVIDIA-A100	Re-LAION-1.4B	1.70e+04	6.79e+04
const	NVIDIA-A100	Re-LAION-1.4B	4.61e+03	1.84e+04
A100 subtotal:			5.56e+05	2.22e+06
NVIDIA H100				
cosine	NVIDIA-H100	DataComp-1.4B	2.09e+04	2.98e+04
cosine	NVIDIA-H100	Re-LAION-1.4B	1.06e+04	1.52e+04
H100 subtotal:			3.15e+04	4.50e+04
Total:			5.87e+05	2.27e+06

Table 4: Total GPU compute and energy consumption for scaling law derivation experiments.

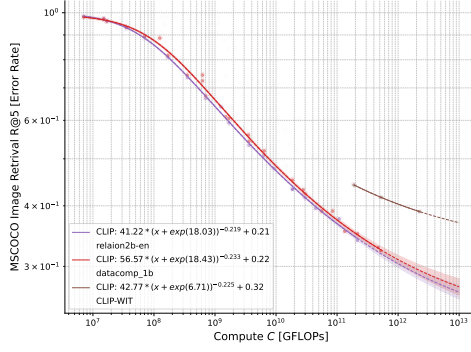


(a) CLIP on WIT, Re-LAION, DataComp

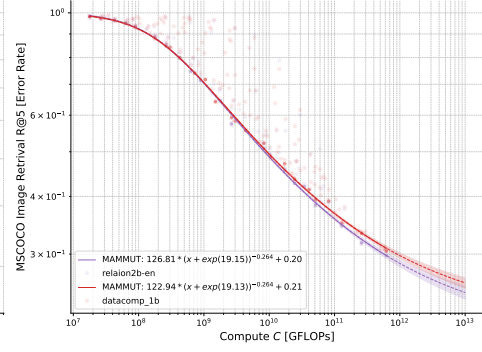


(b) MaMMUT on Re-LAION, DataComp

Figure 7: Scaling laws for IN1k 0-shot performance of openCLIP (left) and openMaMMUT (right), comparing training on DataComp-1.4B and Re-LAION-1.4B. For CLIP we have 3 additional points for OpenAI CLIP [7] models trained on WIT-400M dataset for reference.

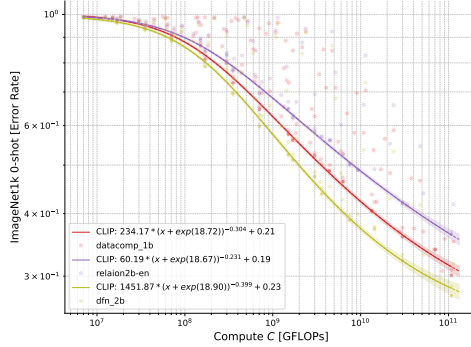


(a) CLIP on WIT, Re-LAION, DataComp

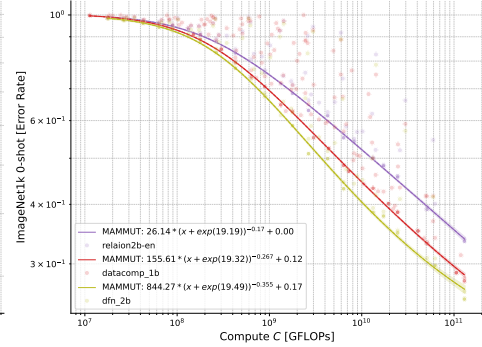


(b) MaMMUT on Re-LAION, DataComp

Figure 8: Scaling laws for MS-COCO image retrieval performance (1- Recall@5) of openCLIP (left) and openMaMMUT (right), comparing training on DataComp-1.4B and Re-LAION-1.4B. For CLIP models we have 3 additional points for OpenAI CLIP [7] trained on WIT-400M dataset for reference.



(a) CLIP on Re-LAION, DataComp, DFN



(b) MaMMUT on Re-LAION, DataComp, DFN

Figure 9: Scaling laws for IN1k 0-shot performance of openCLIP (left) and openMaMMUT (right), comparing training on Re-LAION-1.4B, DataComp-1.4B and DFN-1.4B. Training on DFN-1.4B results in superior performance across scales consistently for both architectures.

Model	ImageNet-1k				MS-COCO Retrieval			
	A_c	B_c	α_c	E_c	A_c	B_c	α_c	E_c
openCLIP	57.862	18.391	-0.227	0.111	53.913	18.413	-0.230	0.216
openMaMMUT	79.970	19.111	-0.233	0.076	119.751	19.122	-0.263	0.212

(a) Fitted scaling law parameters (A_c, B_c, α_c, E_c) for error rate on 0-shot ImageNet-1k classification and MS-COCO retrieval tasks, rounded to three decimal places for models trained on DataComp-1.4B.

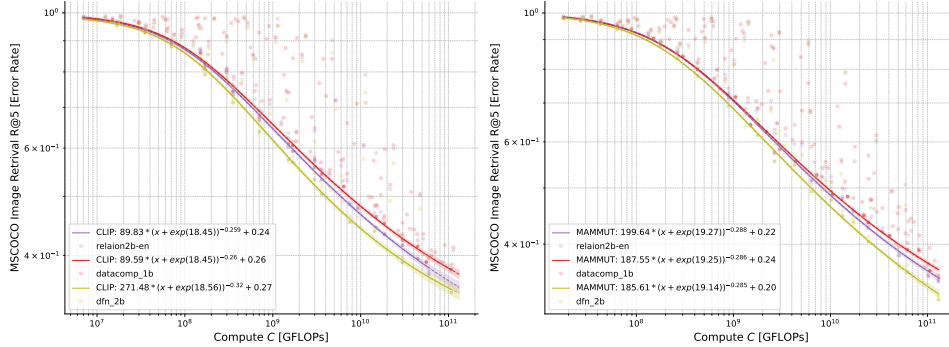
C_0 GFLOPs	IN-1k Err. Rate	$ d\mathcal{L}(C_0)/dC $	COCO R@5 Err. Rate	$ d\mathcal{L}(C_0)/dC $
CLIP				
5.00e+10	9.85e-13		8.44e-13	
1.00e+11	4.21e-13		3.60e-13	
5.00e+11	5.86e-14		4.95e-14	
Average: IN-1k: 4.882e-13, COCO: 4.177e-13				
MaMMUT				
5.00e+10	1.17e-12		9.65e-13	
1.00e+11	4.92e-13		4.03e-13	
5.00e+11	6.54e-14		5.28e-14	
Average: IN-1k: 5.758e-13, COCO: 4.702e-13				

(b) Numerical values of derivatives of fitted functions with respect to compute, in points $5 \cdot 10^{10}$, $1 \cdot 10^{11}$, $5 \cdot 10^{11}$ GFLOPs for both ImageNet-1k error rate and COCO retrieval error rate (1-R@5). MaMMUT consistently exhibits higher values of $|d\mathcal{L}(C_0)/dC|$ which corresponds to higher decrease rate and stronger scalability.

Table 5: Estimated parameters for main scaling law fits for 0-shot ImageNet-1k classification and MS-COCO retrieval, used for openCLIP and openMaMMUT comparison in Fig. 1

Model	ImageNet-1k				MS-COCO Retrieval			
	A_c	B_c	α_c	E_c	A_c	B_c	α_c	E_c
openCLIP	14.769	16.725	-0.168	0.121	6.686	16.209	-0.123	0.089
openMaMMUT	1850.286	20.521	-0.379	0.198	634.190	20.256	-0.335	0.249

Table 6: Fitted scaling law parameters (A_c, B_c, α_c, E_c) for error rate on 0-shot ImageNet-1k classification and MS-COCO retrieval tasks, rounded to three decimal places for models trained on DataComp-1.4B with constant learning rate scheduler.



(a) CLIP on Re-LAION, DataComp, DFN

(b) MaMMUT on Re-LAION, DataComp, DFN

Figure 10: Scaling laws for MS-COCO image retrieval performance (1- Recall@5) of openCLIP (left) and openMaMMUT (right), comparing training on Re-LAION-1.4B, DataComp-1.4B and DFN-1.4B. Training on DFN-1.4B results again in superior performance across scales consistently for both architectures.

Training trial-to-trial variance. To perform trial-to-trial variance sanity check for model pre-training, ensuring that trial-to-trial variance of same runs is substantially smaller than variance due to scaling or hyperparameter tuning, we show downstream task performance on zero-shot IN1K as observed for 3 training runs of the same configuration for reference scales B-32 and B-16 on 640M samples, using same hyperparameters that correspond to minimum loss obtained in tuning, residing on Pareto front. As the results in Tab. 7 suggest, variance is negligibly small compared to difference

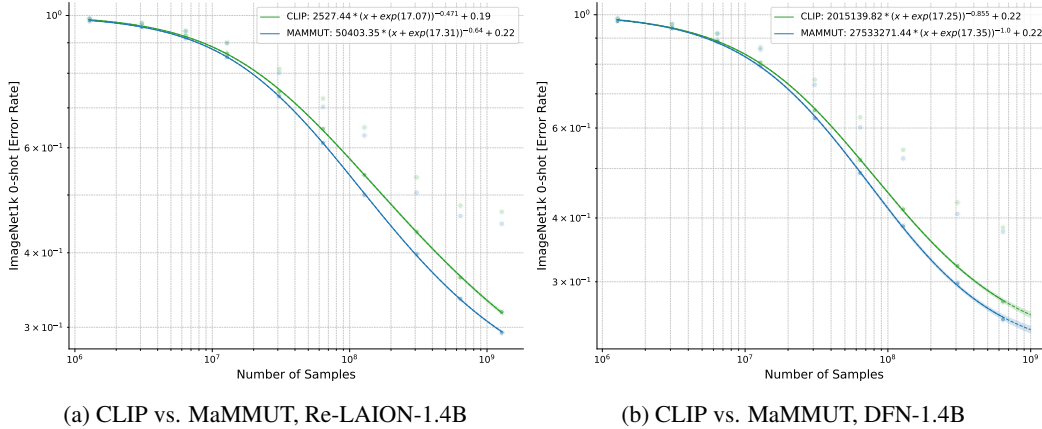


Figure 11: Comparison of data efficiency for CLIP and MaMMUT via scaling laws for IN1k 0-shot classification error on Re-LAION-1.4B (a) and DFN-1.4B (b). MaMMUT is consistently more data efficient on both datasets, which is also in accord with observations from DataComp-1.4B.

due to changing the compute scale. This allows us to conclude that measurements we use for scaling laws derivation can be distorted only insignificantly by trial-to-trial training variance, and scaling trends we observe are valid and shaped dominantly by training compute.

Trial/Training	1	2	3	$\mu \pm \sigma$
B-32 640M	0.58522	0.57866	0.58014	0.58134 ± 0.0034407
B-16 640M	0.66926	0.6668	0.6631	0.666387 ± 0.00310073
L-14 640M	0.72368	0.7203	0.72398	0.722653 ± 0.00204356

Table 7: Trial-to-trial variance control experiment. IN1k zero-shot top-1 on DataComp-1.4B, 640M samples seen. Mean μ and standard deviation σ computed for each reference scale over 3 different training runs. Hyperparameters for each reference scale training run are fixed and correspond to hyperparameters tuned to obtain minimal loss via multiple sweeps for each given reference scale. Trial to trial variance is negligible small compared to performance difference across the scales and is decreasing with increasing performance level.

B.1 Effect of the number of points used for the scaling law fits

For scaling law derivation on DataComp-1.4B (Fig. 1), we used different number of points for MaMMUT (1010 points) and CLIP (672 points). Since MaMMUT architecture was never trained before on open data like DataComp, we had to perform more rigorous hyper-parameter search than for CLIP, hence we ended up with a larger set of measurements. To understand whether the number of measurements that we have obtained affects our conclusions we conduct additional experiment to double check whether there is any difference in the obtained scaling law if working with same number of points for MaMMUT as for CLIP on DataComp-1.4B. We perform bootstrapping, sampling randomly 672 points from 1010 available points for MaMMUT, doing 10 trials, fitting scaling law for each trial and averaging the obtained scaling law coefficients. We observe that the obtained fit coefficients have no significant difference from scaling law obtained with 1000 points 8. Thus, the

Model	A_c	B_c	α_c	E_c
CLIP	57.862083	18.391321	-0.226604	0.111169
MaMMUT (full points)	125.356572	19.289384	-0.255670	0.101112
MaMMUT (same points num. as CLIP)	125.267163	19.301461	-0.255934	0.108208

Table 8: Comparison of different models with their corresponding parameters.

comparison on reduced points shows similar trends. Measurements are balanced for Re-LAION (750 vs 703 points), as well as for DFN (737 vs 732 points).

C Evaluating scaling law fit quality

To validate our scaling law fits, we use a threshold $C_{\text{threshold}}$ to up which we take the data for the fit. We compute RMSE for the held-out points to get a measure of how good each fit is. We compare two $C_{\text{threshold}}$ values (see Tab. 9 and Fig. 20 for DataComp-1.4B dataset). We see that both RMSE and uncertainty (the width of the confidence intervals) decreases as we take more the more points for the fit.

We also compare different functional forms that can be used to fit the data: model with double saturation ($\mathcal{L}(C) = A_c \cdot (C + B_c)^{-\alpha_c} + E_c$) and without a term for irreducible error E_c :

$$\mathcal{L}(C) = A_c \cdot (C + B_c)^{-\alpha_c} \quad (3)$$

We choose first $C_{\text{threshold}} = 2.5 \cdot 10^{11}$ GFLOPs and the second $C_{\text{threshold}} = 5 \cdot 10^{11}$ GFLOPs. As we see from Tab. 9 and Tab. 10 for both values of $C_{\text{threshold}}$ double saturation form (Eq. 1) has consistently lower RMSE than the function without irreducible error. RMSE on held out points provides thus a way to select among various scaling law fits the candidate that provides better prediction accuracy for unseen scales, which in our case is the fit obtained via double saturation functional form (Eq. 1).

We see that the same trend of reducing confidence intervals and thus reducing uncertainty of the predictions when taking more points for the scaling law fit holds also for other tasks like MS-COCO image retrieval and other pre-training dataset Re-LAION-1.4B (see Fig. 20 and Fig. 13 for comparison between ImageNet-1k classification and MS-COCO retrieval and Figs. 18, 19 for Re-LAION-1.4B).

When comparing predictions with actually measured downstream task performance, we see that accuracy for the held-out points is high (Tab. 9). For instance, we measure for 3B samples seen scale on held-out points for openMaMMUT ViT L-14 zero-shot IN1K 0.784, with prediction 0.777 and 95% confidence interval (0.771, 0.783), and for openMaMMUT ViT H-14 0.795, with prediction 0.801 and 95% CI of (0.793, 0.809). Similar accuracy is observed for openCLIP, with actual measurements falling within predicted confidence intervals. **The derived scaling laws provide thus solid ground for comparison on unseen scales that have low amount of repetitions** (less than 3x in case of 3B samples seen scale when training on DataComp-1.4B or Re-LAION-1.4).

As already discussed in Sec. 3.5, the prediction for performance of MaMMUT L-14 on the larger 12.8B samples seen scale (Tab. 1, zero-shot IN1K 0.820, 95% CI (0.815, 0.826)) is therefore only made for low repetition scenario, and to validate it, dataset size larger than currently used 1.4B samples (which gives around 9x repetitions for 12.8B samples seen scale) is required. The measured 0.803 for openMaMMUT L-14 on 12.8B (Tab. 3) is thus expectedly below the prediction, as performance is diminished due to high amount of repetitions, in line with observations by previous works [41, 49].

Model	Samples Seen	GFLOPs	IN1k 0-shot acc	Predicted IN1k 0-shot acc (95% CI)	Predicted (more points) IN1k 0-shot acc (95% CI)
CLIP					
ViT-L-16	3.07e+9	4.07e+11	0.761	0.747 (0.738, 0.755)	–
ViT-L-14	3.07e+9	5.18e+11	0.766	0.753 (0.744, 0.762)	0.759 (0.751, 0.766)
ViT-H-14	3.07e+9	1.14e+12	0.784	0.773 (0.761, 0.784)	0.779 (0.770, 0.789)
<i>RMSE: 1.26e-02 RMSE (more points): 5.90e-03</i>					
MaMMUT					
mammut-ViT-L-14	1.28e+9	2.59e+11	0.749	0.743 (0.737, 0.748)	–
mammut-ViT-L-14	3.07e+9	6.22e+11	0.784	0.773 (0.765, 0.781)	0.777 (0.771, 0.783)
mammut-ViT-H-14	3.07e+9	1.43e+12	0.798	0.797 (0.787, 0.807)	0.801 (0.793, 0.809)
<i>RMSE: 7.57e-03 RMSE (more points): 7.57e-03</i>					

Table 9: Predicting held-out points on compute-optimal Pareto front based on scaling law derivation for the functional form with double saturation (Eq. 1). To check prediction accuracy when extrapolating beyond points taken for the fit, we predict starting from different compute threshold values of $C_{\text{threshold}}^{\text{CLIP}} = 4.07 \cdot 10^{11}$, $C_{\text{threshold}}^{\text{MaMMUT}} = 2.59 \cdot 10^{11}$. $C_{\text{threshold}}$ points themselves are predicted by taking smaller $C_{\text{cutoff}} = 2.5 \cdot 10^{11}$. The last column contains updated predictions made after taking additional data points up to $C_{\text{threshold}}$, showing predictions that extrapolate 2.4 and 5.5 compute factor beyond the fit for MaMMUT, and 1.3 and 2.8 for CLIP. Both confidence interval and RMSE decrease as we take more points. RMSE is consistently lower than RMSE measured for functional form without irreducible error (Tab. 10).

Model	Samples Seen	GFLOPs	IN1k 0-shot acc	Predicted IN1k 0-shot acc (95% CI)	Predicted (more points) IN1k 0-shot acc (95% CI)
CLIP					
ViT-L-16	3.07e+9	4.07e+11	0.761	0.769 (0.764, 0.773)	–
ViT-L-14	3.07e+9	5.18e+11	0.766	0.778 (0.774, 0.783)	0.777 (0.773, 0.782)
ViT-H-14	3.07e+9	1.14e+12	0.784	0.806 (0.802, 0.811)	0.805 (0.801, 0.809)
RMSE: 1.55e-02		RMSE (more points): 1.72e-02			
MaMMUT					
mammut-ViT-L-14	1.28e+9	2.59e+11	0.749	0.757 (0.754, 0.760)	–
mammut-ViT-L-14	3.07e+9	6.22e+11	0.784	0.795 (0.792, 0.798)	0.794 (0.791, 0.796)
mammut-ViT-H-14	3.07e+9	1.43e+12	0.794	0.825 (0.822, 0.828)	0.824 (0.822, 0.827)
RMSE: 1.98e-02		RMSE (more points): 2.26e-02			

Table 10: Predicting held-out points on compute-optimal Pareto front based on scaling law derivation for the functional form without irreducible error (Eq. 3). Comparing prediction quality to the functional form with double saturation (Tab. 9), using same values for C_{cutoff} and $C_{\text{threshold}}$. The last column contains updated predictions made after taking additional data points up to $C_{\text{threshold}}$. Both confidence interval and RMSE decrease as we take more points. RMSE is consistently higher than RMSE measured for functional form with double saturation that includes irreducible error (Tab. 9).

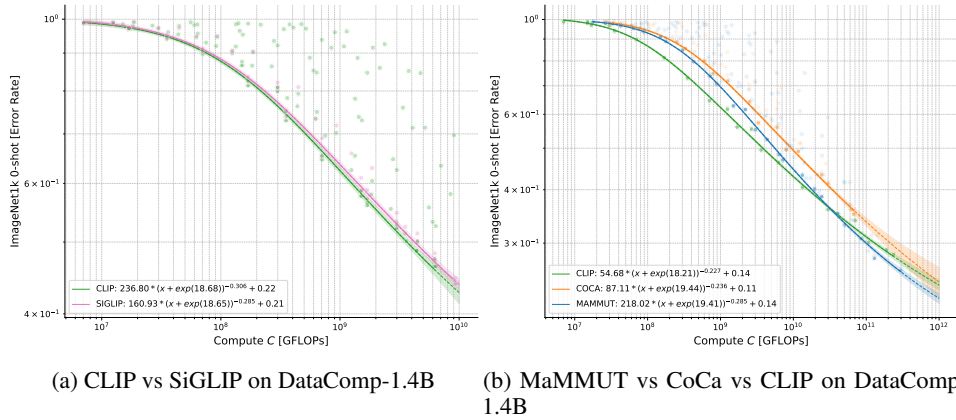


Figure 12: Scaling laws for ImageNet-1k 0-shot classification, comparing SigLIP (left) and CoCa (right) with standard CLIP and MaMMUT using open DataComp-1.4B dataset. SigLIP shows no benefit over standard CLIP, contrary to claims in previous work. CoCa is predicted to be less scalable than MaMMUT, while crossing CLIP is possible, although it is not clear due to high uncertainty for CoCa estimates on larger scales, as measurements on smaller scales for CoCa are not dense enough.

D Datasets comparison

Scaling laws can be also be used as a tool for dataset comparison. Here, we compare performance of models trained on two reference datasets (DataComp-1.4B and Re-LAION-1.4B) for both model architectures – CLIP and MaMMUT, for two downstream tasks – ImageNet-1k 0-shot classification and MS-COCO retrieval. For CLIP, we additionally plot OpenAI CLIP models’ performance that were trained on the WIT-400M dataset. As we see from Fig. 7, **for both CLIP and MaMMUT, training on DataComp-1.4B provides superior scalability for zero-shot ImageNet-1k classification**, compared to training on Re-LAION-1.4B. At the same time, training either on Re-LAION-1.4B or DataComp-1.4B leads to similar scalability and performance on MS-COCO retrieval, with Re-LAION-1.4B being for retrieval slightly more beneficial (Fig. 8).

Using much denser measurements for scaling law derivation, we can also confirm findings from previous work [10], which showed that the closed dataset WIT-400M[7] has better scaling trend on zero-shot classification, but worse scaling trend on zero-shot retrieval when compared to LAION-2B. We observe the same for Re-LAION-1.4B, which is a safety update of LAION-2B used in [10], otherwise being same dataset with less samples due to link rot [21]. This provides further evidence for robustness of scaling law based comparison, showing consistent trends despite major difference in scaling law derivation. Previous work [10] used few samples seen scales of 3B, 12.8B and 34B, which results in high repetition given that only 2B unique samples are contained in LAION-2B [50],

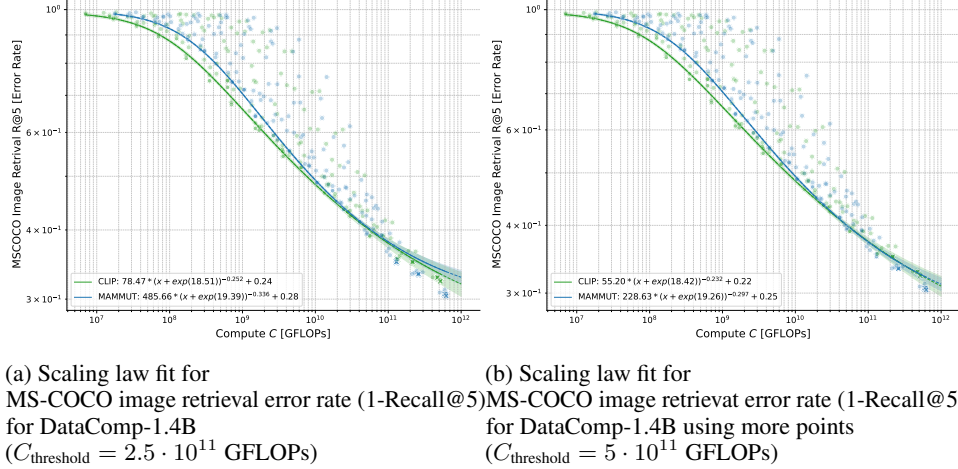


Figure 13: Comparison of the fit quality for MS-COCO image retrieval error rate for openMaMMUT and openCLIP. Adding more points in (b) reduces the uncertainty of the fit, indicated by the width of bands around each curve.

while our work used denser lower samples seen scales up to 3B, doing derivation with unique samples or low repetition only. Despite these differences, derived scaling laws agree in the dataset comparison for each downstream task, predicting same scaling trends in favor of WIT-400M on classification and in favor of Re-LAIION-1.4B on retrieval. DataComp-1.4B can be seen in this comparison as an improved version of Re-LAIION-1.4B, with stronger scalability on classification (Fig. 7) that matches WIT-400M, while obtaining performance for retrieval (Fig. 8) that matches Re-LAIION-1.4B, outperforming WIT-400M.

E Scaling behavior of other architectures

We investigate additional model architectures: SigLIP [18] (CLIP with the sigmoid loss instead of softmax), CoCa [16] and Cap [45] (pure captioner). We train these models on DataComp-1.4B in order to compare with openCLIP and openMaMMUT. Fig. 12 shows the fitted lines for these models. We see that CLIP and SigLIP have very similar scaling behavior on ImageNet-1k classification (Fig. 12 (a)) while openMaMMUT consistently overtakes CoCa on the same compute scale Fig. 12 (b). Notably, our analysis shows that **SigLIP has similar or even worse scalability than CLIP** which contradicts recent claims of SigLIP being a better choice for a vision encoder [18, 14] due to its architectural advantages. Thus, when properly controlling for same training data in our experiments, no benefits for SigLIP can be derived from the obtained scaling law trends. We also observe that text **decoder-only MaMMUT overtakes encoder-decoder CoCa on the same compute scale**, indicating that simpler and more parameter efficient architecture of MaMMUT might be preferable.

Moreover, we see (Fig. 21) that MaMMUT has superior scaling compared to Cap, showing that **combination of contrastive and captioning losses is advantageous**. We see Cap also underperforming standard CLIP, hinting that Cap as captioner only based architecture is not a good candidate for strong scalability in 0-shot regimes, making another case for contrastive losses being important part of scalable architectures for 0-shot classification. It is further important to note that Cap can use only log-likelihood based evaluation for zero-shot classification task, as opposed to CLIP and MaMMUT that in addition can use embedding similarity based evaluation thanks to their contrastive loss. As evident from Fig. 21, embedding similarity based evaluation used in openCLIP and openMaMMUT has strong advantage over log-likelihood based one. It is in addition also much cheaper in execution. Cap has thus architectural disadvantage in not being able to use similarity based evaluation due to missing contrastive loss, which leads to inferior performance in 0-shot regime.

For both comparisons, we see uncertainty getting high when extrapolating to larger scales, which makes it for instance hard to predict whether CoCa might still cross CLIP or not. To reduce uncertainty, it is thus important to both conduct dense measurements at smaller scales and not to cut off measurements at scales too small to be used for proper extrapolation.

Hyperparameter	Value
Model Architecture	mammut-ViT-L-14
Samples Seen	12.8B
Warmup Steps	6000
Global Batch Size	180,224
Learning Rate	2.5×10^{-3}
GPU Hours	3.53×10^4
Number of NVIDIA A100 GPUs	1024

Table 11: Training hyperparameters for openMaMMUT-L-14.

Model	Samples Seen	Warmup	Global Batch Size	Learning Rate
mammut-ViT-S-32	1.28e+06	1000	512	1.00e-03
mammut-ViT-S-32	1.28e+06	1500	512	5.00e-04
mammut-ViT-S-16	1.28e+06	1000	512	5.00e-04
mammut-ViT-S-32	3.07e+06	4000	512	5.00e-04
mammut-ViT-S-16	3.07e+06	4000	512	1.00e-03
mammut-ViT-S-32	6.40e+06	4000	1024	1.00e-03
mammut-ViT-S-32	1.28e+07	4000	2048	2.00e-03
mammut-ViT-S-16	1.28e+07	3000	2048	2.00e-03
mammut-ViT-S-32	3.07e+07	4000	4096	2.00e-03
mammut-ViT-S-16	3.07e+07	3000	4096	2.00e-03
mammut-ViT-S-32	6.40e+07	4000	4096	2.00e-03
mammut-ViT-S-16	6.40e+07	4000	4096	1.50e-03
mammut-ViT-S-32	1.28e+08	4000	8192	2.00e-03
mammut-ViT-S-14	1.28e+08	4000	8192	2.00e-03
mammut-ViT-M-16	1.28e+08	4000	8192	2.00e-03
mammut-ViT-S-14	3.07e+08	4000	16384	2.00e-03
mammut-ViT-M-16	3.07e+08	4000	16384	2.00e-03
mammut-ViT-S-14	6.40e+08	4000	16384	1.50e-03
mammut-ViT-B-16	3.07e+08	4000	16384	2.00e-03
mammut-ViT-B-32	1.28e+09	4000	16384	2.00e-03
mammut-ViT-B-16	6.40e+08	4000	32768	2.00e-03
mammut-ViT-B-14	1.28e+09	4000	90624	2.00e-03
mammut-ViT-L-16	6.40e+08	6000	45056	2.00e-03
mammut-ViT-L-14	6.40e+08	6000	45056	2.00e-03
mammut-ViT-L-14	1.28e+09	4000	90624	2.00e-03
mammut-ViT-L-16	3.07e+09	4000	91136	2.00e-03
mammut-ViT-L-14	3.07e+09	4000	91136	2.00e-03

Table 12: Hyperparameters for MaMMUT models trained on DataComp-1.4B that are located on the Pareto frontier

F Results on additional benchmarks

We also fit scaling laws on the data for other downstream tasks. In the Fig. 14 we show the scaling behavior on DataComp eval suite, which is constituted by averaging over 35 classification tasks from DataComp (see Tab.15 from [19]). Additionally, we provide scaling law fits for ImageNet-V2 and full ImageNet robustness set 0-shot classification performance for both openMaMMUT and openCLIP (Fig. 15). For all of these tasks we see the same trend - openMaMMUT is stronger scalable than openCLIP and has higher performance given the same compute at larger compute scales. This is also valid for the important robustness metrics that reflects out-of-distribution generalization (Fig. 15) - openMaMMUT shows stronger scalable robustness and outperforms openCLIP in robustness at larger compute scales.

Model	Samples Seen	Warmup	Global Batch Size	Learning Rate
ViT-S-32	1.28e+06	1500	512	5.00e-04
ViT-S-16	1.28e+06	1500	512	5.00e-04
ViT-S-16	1.28e+06	1500	512	2.00e-03
ViT-S-32	3.07e+06	1500	1024	5.00e-04
ViT-S-32	6.40e+06	4000	1024	1.00e-03
ViT-S-32	1.28e+07	4000	2048	1.00e-03
ViT-M-32	1.28e+07	3000	2048	1.00e-03
ViT-S-32	3.07e+07	4000	4096	2.00e-03
ViT-S-32	6.40e+07	4000	4096	2.00e-03
ViT-M-32	6.40e+07	10000	4096	1.00e-03
ViT-S-32	1.28e+08	6000	8192	2.00e-03
ViT-S-16	1.28e+08	6000	8192	2.00e-03
ViT-S-32	3.07e+08	8000	16384	2.00e-03
ViT-S-32	6.40e+08	4000	16384	2.00e-03
ViT-S-14	3.07e+08	4000	16384	2.00e-03
ViT-M-32	6.40e+08	6000	32800	2.00e-03
ViT-B-32	1.28e+09	15000	16384	1.00e-03
ViT-L-32	6.40e+08	4000	45056	2.00e-03
ViT-B-16-text-plus	6.40e+08	6000	32768	2.00e-03
ViT-L-32	1.28e+09	4000	90624	4.00e-03
ViT-L-16	6.40e+08	4000	45056	2.00e-03
ViT-L-32	3.07e+09	4000	91136	4.00e-03
ViT-L-14	1.28e+09	4000	90624	4.00e-03
ViT-L-16	3.07e+09	4000	91136	4.00e-03
ViT-L-14	3.07e+09	4000	91136	4.00e-03

Table 13: Hyperparameters for CLIP models trained on DataComp-1.4B that are located on the Pareto frontier.

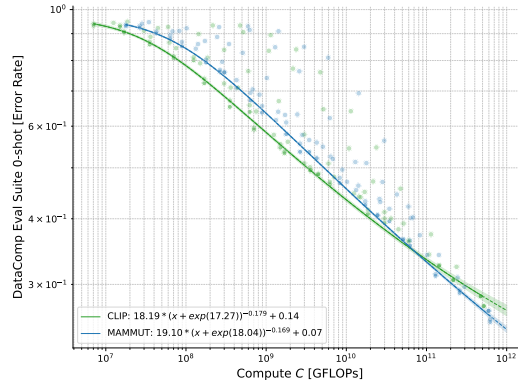


Figure 14: Scaling law on DataComp evaluation suite (average over 35 tasks, 0-shot classification), openCLIP vs. openMAMMUT comparison on DataComp-1.4B

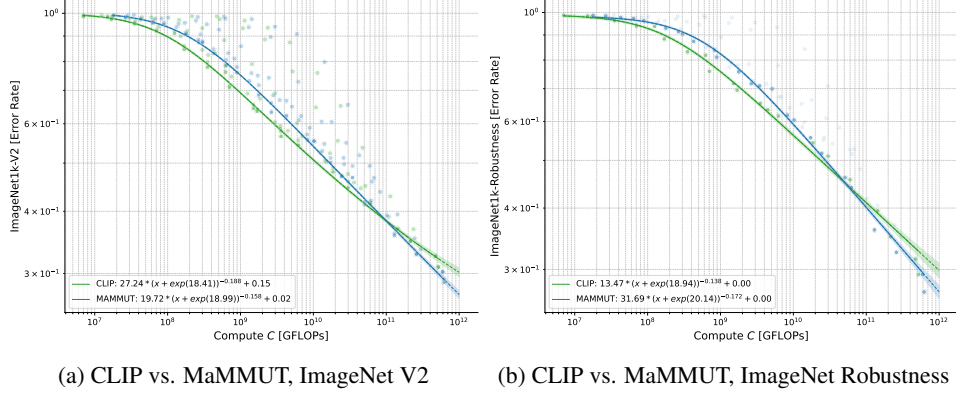


Figure 15: Scaling laws for ImageNet-v2 (left) and ImageNet robustness set (right, averaged performance across 5 datasets ImageNet-v2[27], ImageNet-R[28], ImageNet-Sketch[30], ObjectNet[31], and ImageNet-A[29]), 0-shot classification for openCLIP and openMaMMUT comparison on DataComp-1.4B

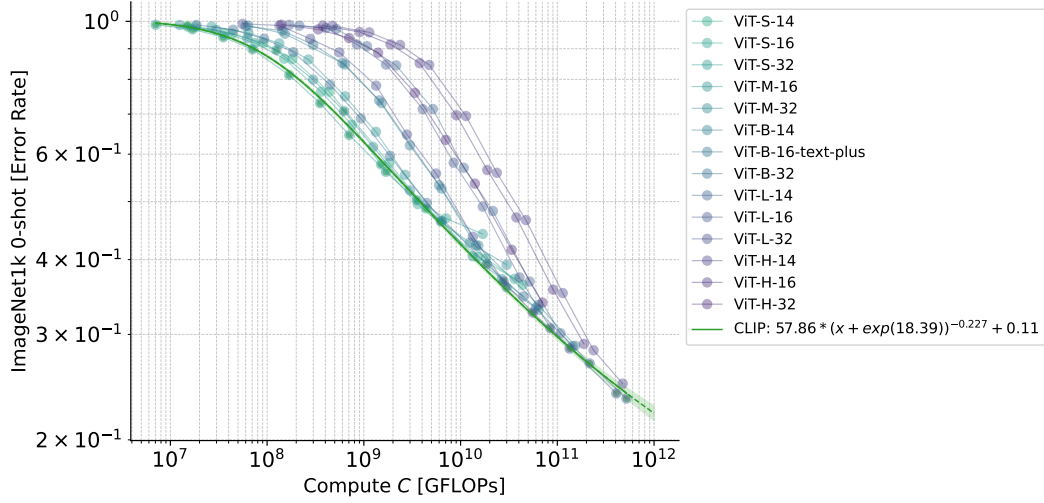


Figure 16: Detailed version of the scaling law fit for ImageNet 0-shot classification error rate for DataComp-1.4B for openCLIP. Cooler colors indicate smaller models. Bigger models are bottlenecked by samples seen scale (require larger samples seen than the smaller ones) and smaller models saturate with increased data and compute scale (over-training regime). Pareto front is composed by taking for each compute budget the points corresponding to models reaching minimal error rate for the given compute. Fit is performed through points on Pareto front.

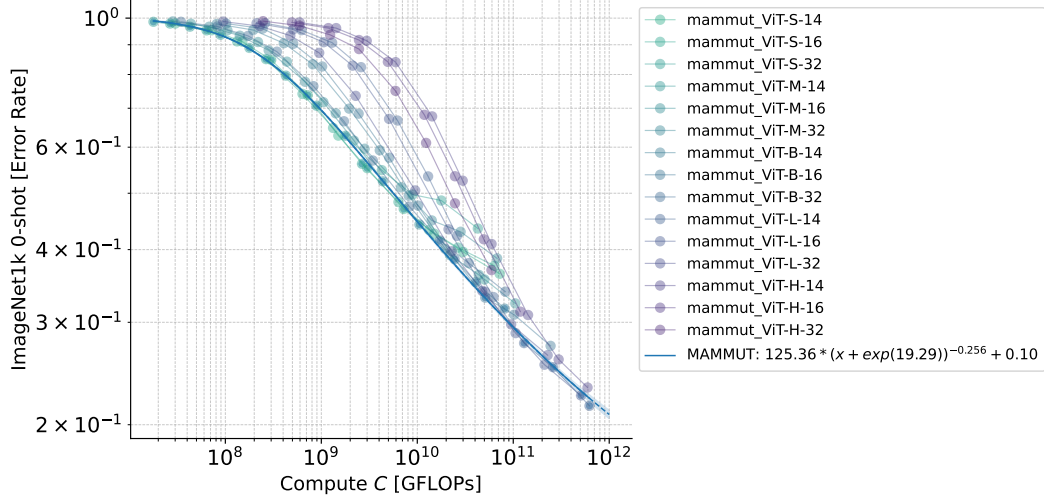


Figure 17: Detailed version of the scaling law fit for ImageNet 0-shot classification error rate for DataComp-1.4B for OpenMaMMUT. Cooler colors indicate smaller models. Bigger models are bottlenecked by samples seen scale (require larger samples seen than the smaller ones) and smaller models saturate with increased data and compute scale (overtraining regime). Pareto front is composed by taking for each compute budget the points corresponding to models reaching minimal error rate for the given compute. Fit is performed through points on Pareto front.

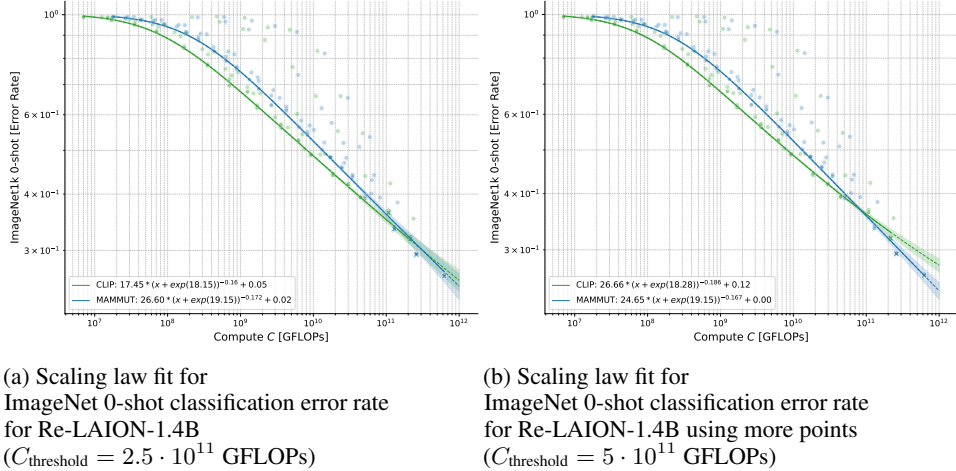


Figure 18: Comparison of the fit quality for ImageNet-1k 0-shot classification error rate for openMaMMUT and openCLIP trained on Re-LAION-1.4B. Adding more points in (b) reduces the uncertainty of the fit compared to (a), indicated by the width of bands around each curve.

Name	Width	Emb	Depth	Params (M)	GFLOPs
ViT-S-32	384/384	384	12/12	63.09	5.51
mammut-ViT-S-32	384/384	384	12/12	85.62	13.91
ViT-S-16	384/384	384	12/12	62.26	11.75
mammut-ViT-S-16	384/384	384	12/12	84.79	20.72
ViT-S-14	384/384	384	12/12	62.21	14.3
mammut-ViT-S-14	384/384	384	12/12	84.74	23.5
ViT-M-32	512/512	512	12/12	103.12	9.74
mammut-ViT-M-32	512/512	512	12/12	134.73	22.1
ViT-M-16	512/512	512	12/12	102.02	20.84
mammut-ViT-M-16	512/512	512	12/12	133.63	34.2
ViT-M-14	512/512	512	12/12	101.95	25.37
mammut-ViT-M-14	512/512	512	12/12	133.57	39.14
ViT-B-32	768/512	512	12/12	151.28	14.54
mammut-ViT-B-32	768/512	512	12/12	183.02	26.91
ViT-B-16	768/512	512	12/12	149.62	39.51
ViT-B-16-text-plus	768/768	768	12/12	210.04	46.78
mammut-ViT-B-16	768/512	512	12/12	290.52	79.7
ViT-B-14	768/512	512	12/12	149.53	49.7
mammut-ViT-B-14	768/512	512	12/12	181.27	63.54
ViT-L-32	1024/768	768	24/12	429.95	43.59
mammut-ViT-L-32	1024/768	768	24/12	510.63	74.28
ViT-L-16	1024/768	768	24/12	427.74	132.37
mammut-ViT-L-16	1024/768	768	24/12	508.42	165.37
ViT-L-14	1024/768	768	24/12	427.62	168.61
mammut-ViT-L-14	1024/768	768	24/12	508.29	202.56
ViT-H-32	1280/1024	1024	32/24	989.02	109.81
mammut-ViT-H-32	1280/1024	1024	32/24	1191.06	192.97
ViT-H-16	1280/1024	1024	32/24	986.26	294.78
mammut-ViT-H-16	1280/1024	1024	32/24	1188.3	385.72
ViT-H-14	1280/1024	1024	32/24	986.11	370.28
mammut-ViT-H-14	1280/1024	1024	32/24	1188.14	464.39

Table 14: Hyper-parameters of architectures we consider. **Width** refers to encoder width, **Emb** refers to embedding size, **Depth** refers to number of layers, **Params** refer to the number of parameters in millions, and **GFLOPs** refer to total GFLOPs per forward pass. Entries in the form of A / B denote image and text parameters respectively. There are more parameters in MaMMUT models because of the additional cross-attention layers.

G Additional training details

In Tab. 11 we provide hyperparameters that were used for training openMaMMUT-L-14. Additionally, in the Tab. 13 and 12 we provide training hyperparameters for all models and sample seen scales that were used for scaling law fits (i.e. models that are located on the Pareto frontier) for openMaMMUT and openCLIP respectively. Tab. 14 we provide overview of model architectures parameters used for training openCLIP and openMammut.

H More details on fine-tuning for segmentation and scaling laws

Following prior work on how to benchmark vision foundation models for semantic segmentation [35], we evaluate CLIP and MaMMUT on semantic segmentation by fine-tuning them end-to-end using a linear decoder on ADE20K [34]. Regardless of the patch size used during pre-training, we interpolate the patch size of all models to 14×14 , to ensure a fair comparison. We use an image input size of 224×224 and thus interpolate the positional embedding to 16×16 . Hyperparameters used for training are consistent with [35], except the use of a linear learning rate warmup of 1500 steps, an epoch-based schedule of 31 epochs, and a batch size of 16 without gradient accumulation, following [36].

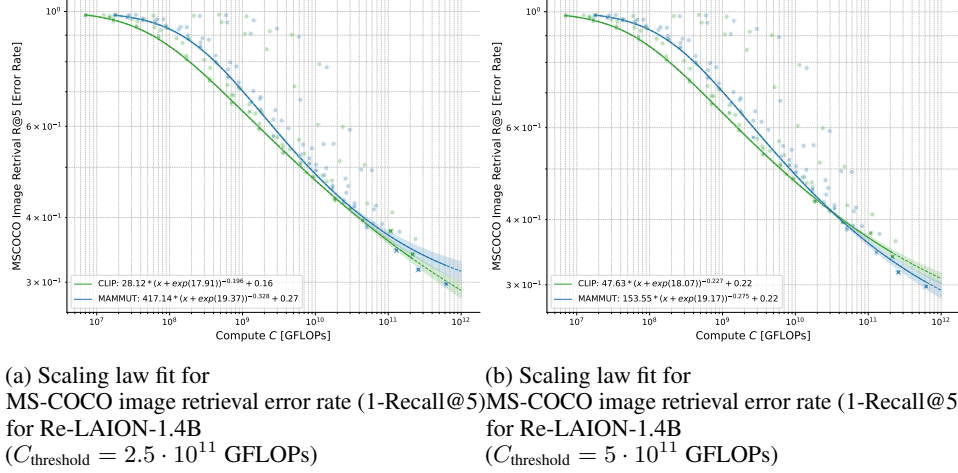


Figure 19: Comparison of the fit quality for MS-COCO image retrieval error rate for openMaMMUT and openCLIP trained on Re-LAION-1.4B. Adding more points in (b) reduces the uncertainty of the fit compared to (a), indicated by the width of bands around each curve.

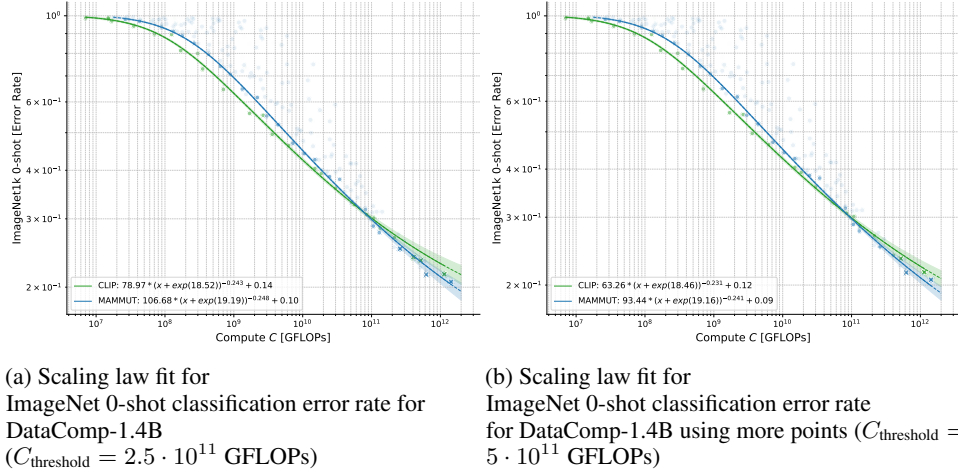


Figure 20: Comparison of the fit quality for ImageNet-1k 0-shot classification error rate for openMaMMUT and openCLIP trained on DataComp-1.4B. Adding more points in (b) reduces the uncertainty of the fit compared to (a), indicated by the width of bands around each curve.

We fine-tune pre-trained models up to and including ViT-L and 3B samples seen, with different pre-training hyperparameters. We evaluate using a sliding window approach, again following [35].

Fig. 22 and Fig. 23 show the fitted scaling laws for CLIP and MaMMUT, respectively. Tab. 15 shows the corresponding estimated scaling law fit parameters.

	A_c	B_c	α_c	E_c
CLIP	18.407549	17.577295	-0.209187	0.468456
MaMMUT	352.152176	18.759619	-0.356718	0.497617

Table 15: Fitted scaling law parameters (A_c, B_c, α_c, E_c) for segmentation error rate.

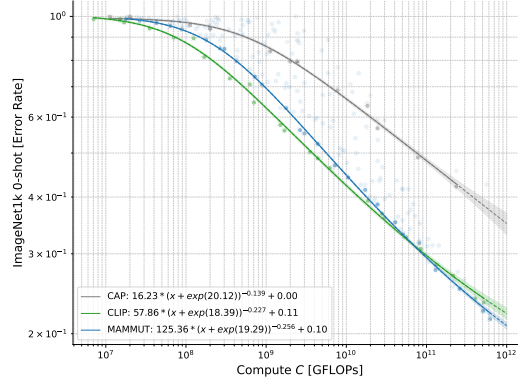


Figure 21: Scaling law fit for ImageNet-1k 0-shot classification, comparing MaMMUT, CLIP and Cap (captioning only). Cap can be only evaluated via log-likelihood, which is more expensive as similarity based evaluation used by CLIP and MaMMUT, as Cap misses contrastive loss in its architecture, which makes it disadvantageous for 0-shot setting.

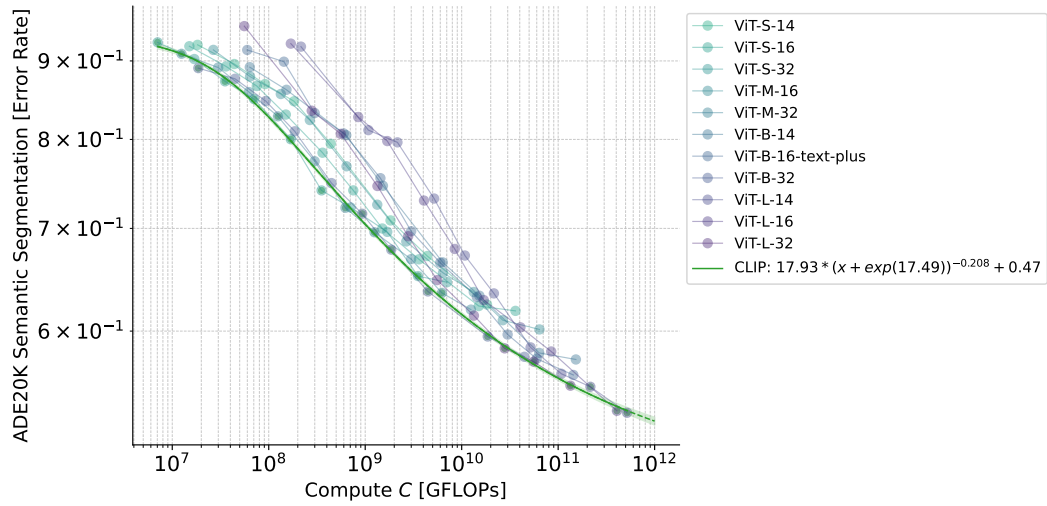


Figure 22: Downstream semantic segmentation performance of CLIP pre-trained on DataComp-1.4B and fine-tuned on ADE20K. Error rate ($1 - \text{mIoU}$).

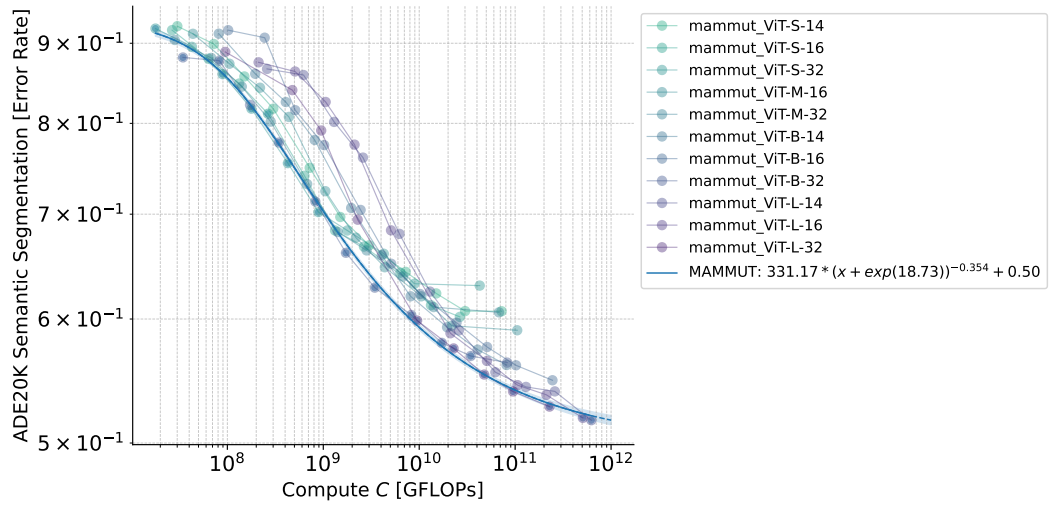


Figure 23: Downstream semantic segmentation performance of MaMMUT pre-trained on DataComp-1.4B and fine-tuned on ADE20K. Error rate ($1 - \text{mIoU}$).

I Author contributions

- **Marianna Nezhurina**: established major part of scaling law fitting procedures. Performed analysis of scaling law fit quality, derived predictions and confidence intervals. Conducted major part of data analysis. Performed initial training experiments with openCLIP and openMaMMUT. Established environments for experiments across various supercomputers. Supported compute resource acquisition. Established infrastructure for distributed dataset acquisition via Ray. Obtained Re-LAION and part of DFN dataset. Co-organized automated experiments data collection and analysis. Wrote the manuscript.
- **Tomer Porian**: co-designed and performed const lr schedule scaling law derivation experiments. Extended automated experiments execution for const lr schedule experiments. Collected and analyzed data, provided further input for scaling law fitting procedures. Co-wrote the manuscript.
- **Tommie Keressies**: fine-tuning experiments for dense prediction segmentation, evaluating segmentation via different modes, scaling law derivation for segmentation, data collection and analysis. Co-wrote the manuscript.
- **Giovanni Puccetti**: openMaMMUT implementation in openCLIP, initial experiments with openCLIP and openMammut training. Initial co-design and implementation of automated experiments execution. Proof reading the manuscript.
- **Romain Beaumont** Re-LAION safety maintenance, hash filtering and re-packaging. Toolsets for dataset download and composition. Proof reading the manuscript.
- **Mehdi Cherti**: led the project, supported compute resource acquisition. Co-established environments for experiments across various supercomputers. Obtained DataComp, DFN and part of Re-LAION dataset. Designed and implemented automated experiments execution and evaluation. Wrote procedure for const lr schedule experiments. Conducted scaling law derivation experiments (DataComp, Re-LAION, DFN; openMammut, openCLIP, Cap). Designed and implemented evaluation. Organized automated experiments data collection and analysis. Collected and analysed the experimental data. Wrote the manuscript.
- **Jenia Jitsev**: led and coordinated the project, acquired compute resources. Organized data transfer (DataComp, Re-LAION) across the supercomputers. Co-established environments for experiments across various supercomputers. Co-designed automated experiments execution. Defined, designed and conducted scaling law derivation experiments (DataComp, Re-LAION, DFN; openMammut, openCLIP, CoCa, SigLIP). Collected and analysed the experimental data. Trained openMammut-L-14 on 12.8B of DataComp-1.4B, following the scaling law predictions. Led manuscript writing, wrote the manuscript.