The truth they do not want you to know about CLIP on 🔥 small(er) image-text datasets 🔥

Vitus Benson Mingxiao Li Beatrice Portelli Gorjan Radevski Emanuele Vivoli

Nicholas Moratelli

Abstract

Simultaneous self-supervised training of vision and language models leads to robust and powerful embeddings. Commonly, large-scale models and datasets are leveraged. However, there is a growing interest in models of medium size, that can be run at scale in production. Here, we study how common strategies to improve training speed and accuracy of contrastive language-image pretraining perform at medium scale. For this, we evaluate patch dropout, sigmoid loss, token merging and unimodal pre-training on the CC3M dataset. Our results partially contradict findings at larger scale: while the sigmoid loss remains valuable at medium scale, patch dropout decreases performance. Unimodal self-supervised pretraining of a vision model directly shows a strong ImageNet zero-shot accuracy, yet language-image pretraining still improves it. Both patch dropout and token merging improve training throughput, while decreasing validation performance.

1. Introduction

Research on Vision-language foundation models [6, 10] demonstrates improved downstream task performance as the dataset and model size increases. However, this induces increased computation cost, where models are trained on datasets ranging up-to 400 million images (e.g., in the case of CLIP) – a process which takes thousands of GPU days. Recently, Li *et al.* [5] propose a CLIP variant that randomly masks and removes a large portion of the image patches during training, resulting in 3.5× faster vision-language pretraining with comparable accuracy. Similarly, Bolya *et al.* [1] propose Token Merging (TokenMerge), which combines similar tokens in a transformer model using a light-weight matching algorithm, leading to 2× the ViT throughput when applied at both training and inference time. Moreover, the image-text contrastive loss (which aligns the image and text embeddings), computes the softmax over the cosine-similarities between all image-text pairs in the batch. To counter, Zhai *et al.* [9] propose the Sigmoid Loss (SigLoss), which eliminates the Softmax over the image-text similarities, and in turn relies *only* on the pairwise computations – which reduces training time. Despite the training efficiency improvements on top of CLIP demonstrated

Figure 1. Author Contributions: Vitus and Beatrice curated the experiments. Mingxiao took care of TokenMerge. Gorjan generated these beatuful lines out of the initial draft of Emanuele, and together they run some initial experiments. Nicholas supervised the whole project.

by these methods, they nevertheless rely on expensive training of large models (e.g., ViT-G) on large-scale datasets such as LAION400M [7], which, in itself, is a cumbersome procedure demanding thousands of GPU hours.

In this paper, we focus on how smaller models (e.g., ViT-S) behave when trained on image-text datasets with orders of magnitude fewer samples, such as CC3M [8]. We perform exhaustive ablation studies and probe different components of the image-text contrastive loss pipeline, where we encounter trends that deviate from findings in the literature. Namely, we observe that on data such as CC3M, FLIP [5] performs worse than the plain CLIP. Secondly, we observe that simply using a self-supervised pre-trained image backbone – DINO [2], and a pre-trained text-encoder – XLM-RoBERTa [3], yields a significantly better sample-efficiency w.r.t. more sophisticated methods. We further observe that the recently proposed SigLIP [9], when combined with pre-trained image and text encoders yields a further improvement in terms of sample-efficiency. Overall, our experiments suggest that reasonable zero-shot classification performance can be achieved by training methods on small-scale datasets.

2. Experimental setup

In total we use four methods for our experiments: the plain CLIP [6], and the FLIP [5], SigLoss CLIP (SigLit) [9] and TokenMerge [1] variants. Compared to CLIP, FLIP uses masking of the image patches in the vision encoder, SigLit uses a different loss function, and TokenMerge performs the merging of similar image patches. Subsequently, our research ventured into optimizing the FLIP architecture with insights derived from [1], which proffered a novel approach merging object parts into singular tokens as a means to enhance model performance. Consequently, our investigative efforts concentrated on the alteration of the loss function. In fact, the original configuration of CLIP comes with the softmax-based contrastive loss. In an attempt to bolster the model's performance, we substitute this loss function with the Sigmoid loss, which was motivated by avoiding pairwise similarities for the purpose of normalization.

The following training strategies are taken into consideration: FLIP, SigLoss, and TokenMerge. They are used in isolation or combined among each other to create a set of six methods. Additionally, we experiment with different initializations of the vision and the text encoder. In the case of a pre-trained vision encoder, we use DINO which performs self-supervised pre-trained on the ImageNet training set[1]. In the case of a pre-trained RoBERTa model, we initialize the weights with XLM-RoBERTa, which is pre-trained using mask language modelling on CommonCrawl[2].

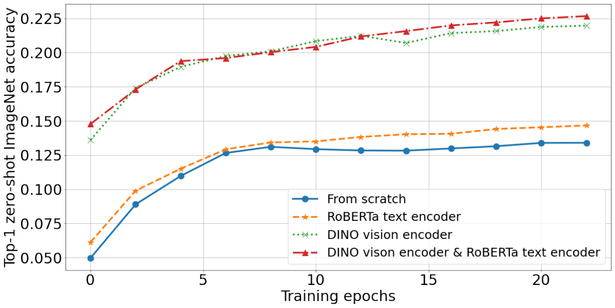

Figure 2. Model performance with differently pre-trained vision and text backbones.

All experiments are performed using 4 GPUs across 2 nodes. We train for 24 epochs with a total batch size of 2048. The employ a cosine annealing learning rate schedule with 2000 linear warm-up steps, and AdamW optimizer. The experiments in which we use FLIP, we use a patch dropout of 0.5. All models are trained on the CC3M dataset. The evaluation is performed by measuring the top-1 zero-shot classification validation accuracy on ImageNet-1K [4]. Our open source code is accessible on GitHub.

3. Experiments and Discussion

Effect of pre-trained vision and text backbones. We report results in Fig. 2. We observe that training from scratch (i.e., CLIP) the vision and text encoders on the image-text data yields the weakest performance. Using a pre-trained RoBERTa encoder improves the performance only moderately, while using a pre-trained image-encoder (DINO) yields a significant performance boost. Lastly, employing both a pre-trained image and text encoder yields the highest performance.

Effect of different training strategies We report results in Fig 3. Interestingly, we observe that FLIP training performs even worse that regular the CLIP, when trained on CC3M, which contradicts the findings in the literature. On the other hand, employing SigLip outperforms all methods we experiment with (yields 9% higher top-1 ImageNet accuracy over CLIP). Additionally, we also experiment with TokenMerge [1], which negatively affects the performance of CLIP, however, as demonstrated next, improves the training sample-efficiency.

[1] huggingface-ViT-dino
[2] huggingface-RoBERTa

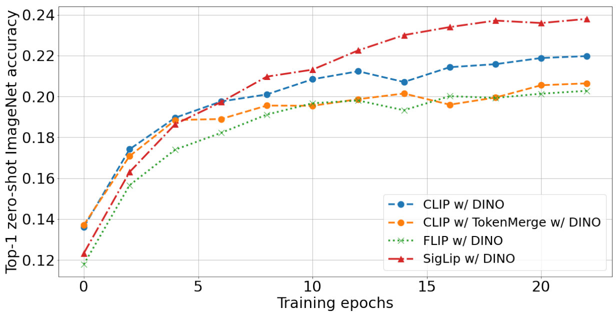

Figure 3. TokenMerge, FLIP and SigLoss accuracy against standard CLIP with pre-trained vision backbone.

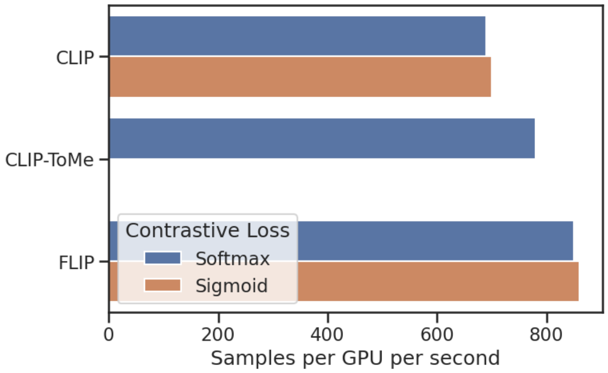

Figure 4. Training throughput, other parameters held equal.

Patch dropout and token merging increase training throughput. Training speed, in our case measured by throughput, that is how many samples are processed per GPU per second, is highest for SigFLIP (Fig. 4). Our token merging implementation has 16% higher throughput than standard CLIP and our FLIP implementation has 23% higher throughput than standard CLIP, all other training parameters held equal. Using SigLip increases both CLIP and FLIP throughput by another 1%.

4. Conclusion

Overall, we uncover that image-text training methods yield promising performance even when trained on datasets orders of magnitude smaller than datasets commonly used in the literature. We observe that among all methods we benchmark, training on image-text paired data using a sigmoid loss, while starting from self-supervised pre-trained image and text encoders is the most sample efficient, i.e., it yields the highest performance when trained for a fixed number of epochs. Finally, methods like token merging and patch dropout can speed up training, but in contrast to the larger scale, they do come at a cost of lower zero-shot accuracy.

Limitations. Notably, we did not perform extensive hyperparameter tuning for each of these methods, but rather, tested all methods with a standard set of hyperparameters. While indeed similar, these methods may behave differently when the optimal set of hyperparameters is chosen for each of them.

Future work may further investigate the behavior of image-text pre-training at medium scale. For instance, the TokenMerge method [1] may be investigated within the text model, the pre-trained backbones may be locked [10] (i.e., LiT way of pre-training) or a dataset of higher quality compared to CC3M (but the same size), can be leveraged for further insights.

References

[1] Daniel Bolya, Cheng-Yang Fu, et al. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 1, 2, 3

[2] Mathilde Caron, Hugo Touvron, et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] Alexis Conneau, Kartikay Khandelwal, et al. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 2

[4] Jia Deng, Wei Dong, et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[5] Yanghao Li, Haoqi Fan, et al. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 1, 2

[6] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[7] Christoph Schuhmann, Richard Vencu, et al. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[8] Piyush Sharma, Nan Ding, et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[9] Xiaohua Zhai, Basil Mustafa, et al. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 1, 2

[10] Xiaohua Zhai, Xiao Wang, et al. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 1, 3