Efficient Multimodal Alignment: To Freeze or Not to Freeze?

Till Aczel
ETH Zürich
taczel@ethz.ch

Roger Wattenhofer
ETH Zürich
wattenhofer@ethz.ch

Abstract

Language-image pretraining creates a joint representation space between the two modalities where images and texts with similar semantic information lay close to each other. Language-image models are often trained from scratch without taking advantage of unimodal pretrained models. By aligning the representation spaces of two modality-specific encoders, our model achieves 74.7% accuracy on the ImagenNet1K validation set, at two orders of magnitude lower training cost. In this work, we highlight the importance of unfreezing the CLS tokens of uni-modal transformer encoders to create a joint embedding space. Freezing the image and text CLS tokens reduces the mean accuracy from 37.5% to 19.4% on the 38 evaluation benchmarks.

1 Introduction

In early 2022, newly emerging language-image models like DALL-E (and later Stable Diffusion, Midjourney, etc.) sent shock-waves through the machine learning community. Training large multimodal models from scratch is computationally expensive, and beyond the capabilities of many small organizations. Instead, what about combining already pretrained unimodal models? This may reduce the computational cost significantly, and open up multimodal models to organizations with limited resources. Model fine-tuning can strike a balance between adapting a model to the given task, while not forgetting the initial representations.

A joint multimodal representation space needs to satisfy two properties. 1) Samples with similar semantic information need to be closer to each other than samples with dissimilar semantic information. 2) The representation spaces of the two modalities need to align. Encoders trained on their respective data modality have learned a representation space that satisfies the first property. Freezing some components of the unimodal encoders helps to keep representations that satisfy property 1. At the same time, unfreezing (fine-tuning) parts of the model helps to learn representations that satisfy property 2. Which parts of the model to freeze/unfreeze plays an important role in the model aligning vs. not-forgetting balance.

In this work, we adapt the Contrastive Language-Image Pretraining (CLIP) [1] framework. Instead of training models from scratch, we align the representation spaces of unimodal encoders. Zhai et al. [2] investigate multimodal representation alignment while freezing the pretrained text or image encoder. We explore freezing/unfreezing components of the transformer encoders, to improve the zero-shot performance at a reduced training cost. The main contributions of this paper are:

1. We demonstrate that by aligning unimodal pre-trained model representations, on a small scale, comparable results can be achieved at a fraction of the computation and data cost. By taking advantage of unimodal encoders, our model achieves 74.7% ImageNet1k [3] accuracy, which is on par with models of comparable size, with ∼99% reduced training cost.

Preprint.

2. We show the importance of unfreezing the patch/token embeddings while aligning the transformer encoder representation space. Surprisingly, fine-tuning the image and text CLS tokens is crucial to aligning language-image representation spaces. By freezing just the two CLS tokens, mean accuracy drops from 37.5% to 19.4% on 38 image classification and image-text retrieval benchmarks [4].

2 Related work

Radford et al. [1] introduce a contrastive learning framework that establishes a unified embedding space for textual and image data. This approach maximizes the similarity for web-scraped image-text pairs and minimizes it for unrelated pairs. The utilization of this extensive and diverse dataset enables these models to understand a wide range of concepts and relationships across modalities. Leveraging the power of their joint embedding space, language-image models are achieving exceptional results in a wide range of tasks, including zero-shot image classification and image-text retrieval [1; 5; 2; 6; 7; 8; 9; 10], but also semantic image generation [11; 12; 13; 14], showcasing their capacity for multimodal understanding.

Most relevant to our work are papers demonstrating that fine-tuning pretrained representation models can both speed up training and improve performance. Sun et al. [8] align an EVA [15] vision transformer with a CLIP [1] and OpenCLIP [7] text encoder to achieve state-of-the-art results with reduced training costs. To mitigate over-fitting while aligning image and text encoders Zhai et al. [2] freeze the image encoder, and only fine-tune the text encoder. Their findings reveal that using a frozen pretrained vision encoder not only speeds up training time but also increases model performance, even for training runs with 20 billion seen samples. In contrast to Zhai et al. [2], we focus on the importance of fine-tuning the input embedding representations, with a special focus on the transformer CLS tokens.

In Section 4, we compare our results against prior works. Pham et al. [5] increase the batch size and training schedule of the CLIP [1] model. As contrastive training relies on the negative samples in the batch, CLIP-style models can benefit from a larger batch size [1; 5; 7]. Cherti et al. [7] release an open-source state-of-the-art language-image model with training code, trained on the LAION5B [16] dataset. To remove the dependence on batch size, Zhai et al. [9] introduce a pairwise sigmoid loss for contrastive language-image pretraining. Randomly dropping some of the transformer input tokens acts as regularization and also increases training speed [10]. Yu et al. [6] combine the contrastive loss with a multimodal autoregressive captioner loss to learn the joint image-text representations. Moreover, Schuhmann et al. [16] released LAION5B, the first large-scale open-source web-scraped image-text dataset. Models trained on less but higher quality data can outperform models that were trained on noisy data [4]. The DataComp [4] challenge focuses on filtering strategies to create a high-quality image-text dataset.

3 Experiments

For all experiments we use a subset of the CommonCrawl [17] image-text pair dataset, we filter the medium version of the DataComp [4] dataset by their *Image-based* $\cap$ *CLIP score (L/14 30%)* filtering strategy. The dataset contains 13 million image text pairs. For evaluation, we use the ImageNet1K [3] accuracy and the mean of 38 image classification and retrieval tasks from the DataComp [4] challenge.

The models are trained with a batch size of 4096 until 12.8 million training samples are seen, which is slightly less than the dataset size. We optimize the symmetric cross-entropy loss [1] between the CLS token of the vision transformer output and the projected CLS token of the text transformer. For a fair comparison, all models use a similar number of parameters in both the vision and language transformers. The vision encoders use the Vit-B/16 architecture, which is a 12-layer transformer, with an input image size of 224, divided up into 16×16 patches. All text embedders use a 12-layer transformer encoder. To speed up training, all models were trained with bfloat16 mixed precision. We adopt the hyperparameters from the small scale DataComp [4] challenge for the remaining model settings.

4 Results

Reduced computational cost

Aligning unimodal representation spaces, rather than starting from scratch reduces the computational cost. This reduction in computational requirements enables researchers, especially those with limited access to computing resources to investigate contrastive language-image pretraining. Notably, this method achieves an approximately 99% reduction in training costs while delivering performance on par with the best performing (EVA-02-CLIP [8]) similarly-sized model, as evidenced by the results in Table 1.

seen ImageNet	# seen samples	model	# parameters image	# parameters text	dataset size	ImageNet acc (%)
✗	52,000M	BASIC-M [5]	168M	184M	6,600M	81.5
✗	40,000M	BASIC-S [5]	25M	108M	6,600M	71.9
✗	34,000M	OpenCLIP [7]	86M	63M	2,000M	70.2
?	13,000M	CLIP [1]	86M	63M	400M	68.3
?	13,000M	FLIP [10]	86M	53M	400M	68.0
?	8,000M	EVA-02-CLIP [8]	86M	63M	2,400M	74.7
✓	18,000M	LiT [2]	86M	110M	4,000M	73.9
✓	9,000M	SigLIP [9]	86M	63M	4,000M	73.4
✓	**128M**	ours	86M	110M	13M	74.7
✓	**13M**	ours	86M	110M	13M	68.3

Table 1: Comparison to previous methods on the ImageNet1K validation set. The number of seen samples is a reasonable measure of computational cost, for models with a similar number of parameters. Our model achieves comparable performance to state-of-the-art models with two orders of magnitude less computational resources and two orders of magnitude smaller datasets. LiT [2], SigLIP [9] and our model fine-tunes vision transformers which were pretrained on the ImageNet training set, and the CLIP [1], FLIP [10] and EVA-02-CLIP [8] image-text training dataset might contain some samples from the evaluation dataset.

Input embedding fine-tuning

To align the unimodal pre-trained representations it is beneficial to freeze some parts of the model and fine-tune the rest. We experiment with freezing and unfreezing the patch embeddings/token embeddings and the transformer blocks. The vision and language CLS tokens are frozen/unfrozen together with the patch embeddings and the token embeddings, if not specified otherwise. For a vision model V and text model T, we note if the patch/token embeddings are frozen/unfrozen in the lower index and if the transformer blocks are frozen/unfrozen in the upper index. From the 16 possible combinations, freezing the vision encoder transformer blocks, while unfreezing the vision patch embeddings and the whole text encoder ($V_u^f T_u^u$) has the best performance. In Table 2 we ablate the freezing/unfreezing of the input embeddings and the transformer blocks, by changing one at a time compared to the best setting $V_u^f T_u^u$.

image pretrain	text pretrain	$V_u^f T_u^u$	$V_f^f T_u^u$	$V_u^f T_f^u$	$V_u^u T_u^u$	$V_u^f T_u^f$
Supr. INet 21, 1k [18]	BERT [19]	**43.0±.2**	39.4±.2	21.5±.1	39.6±.1	36.7±.0
Supr. INet 21, 1k [18]	mBERT [19]	**42.7±.6**	39.3±.2	17.3±.1	40.2±.4	37.2±.4
Supr. INet 21, 1k [18]	RoBERTa [20]	**41.2±.5**	37.9±.1	23.9±.3	38.4±.4	31.5±.1
Supr. INet 21, 1k [18]	T5 [21]	**35.7±.4**	32.2±.5	20.0±.2	34.9±.3	20.3±.0
Supr. INet 21k [18]	BERT [19]	**41.5±.9**	34.5±.5	20.6±.1	40.2±.4	34.4±.1
DINO INet 1k [22]	BERT [19]	**37.2±.1**	**37.3±.2**	18.2±.2	36.4±.3	31.0±.2

Table 2: Mean accuracy on the 38 benchmark tasks over different unimodal representation space alignments. Each setting was run with 3 seeds, mean and standard deviation are reported.

In the case of all text encoders, the most significant decline in performance is observed when freezing the text tokenizer ($V_u^f T_f^u$). Intriguingly, this decline cannot be attributed solely to the number of parameters within the text tokenizer. For all encoders, the tokenizer has fewer parameters than the transformer, see Table 3. The joint representation space between text and images requires information that is not present in the tokenizer embeddings pretrained on the unimodal pretext tasks. Fine-tuning the input embeddings allows the model to re-learn the required information.

model	# of params.	tokenizer	# of tokens	tokenizer # of params.
BERT [19]	110M	WordPiece [23]	30k	23M
mBERT [19]	110M	WordPiece [23]	110k	84M
RoBERTa [20]	125M	BPE [24]	50k	38M
T5 [21]	220M	SentencePiece [25]	32k	24M

Table 3: All text encoder transformers have more parameters than their tokenizer, but freezing the text encoder transformer blocks leads to a smaller performance drop than freezing the tokenizer.

In both the vision and language transformers the information is aggregated via the CLS token. We are comparing the effect of unfreezing just the CLS token embeddings against unfreezing both the CLS token and patch/token embeddings. For all experiments, the Imagenet 1K vision transformer [18] and BERT text encoder [19] are aligned, and evaluated on the 38 dataset benchmark [4]. Starting from the optimal configuration ($V_u^f T_u^u$), freezing the image patch embeddings and the non-CLS token text embeddings, the performance experiences a decline from 43.0% to 37.5%. When also freezing the image and text CLS embeddings performance drops further from 37.5% to 19.4%. Fine-tuning of the CLS token embeddings is needed, as freezing just 2×768 parameters the model's accuracy experiences a drop to roughly half of its original performance.

Aligning unimodal representations might converge faster but to a lower plateau. In Figure 4, it can be observed that accuracy improvement starts to slow down after just a few million seen samples. Zhai et al. [2] results reveal that freezing the vision encoder while unfreezing the text encoder (in our paper, we call this $V_f^f T_u^u$) performs better than unfreezing everything (in our notation $V_u^u T_u^u$), even for longer training runs. The finding that unfreezing everything results in worse performance than freezing some parts of the model suggests that the unimodal encoders have learned representations that are challenging, or even impossible to learn solely through the contrastive training objective.

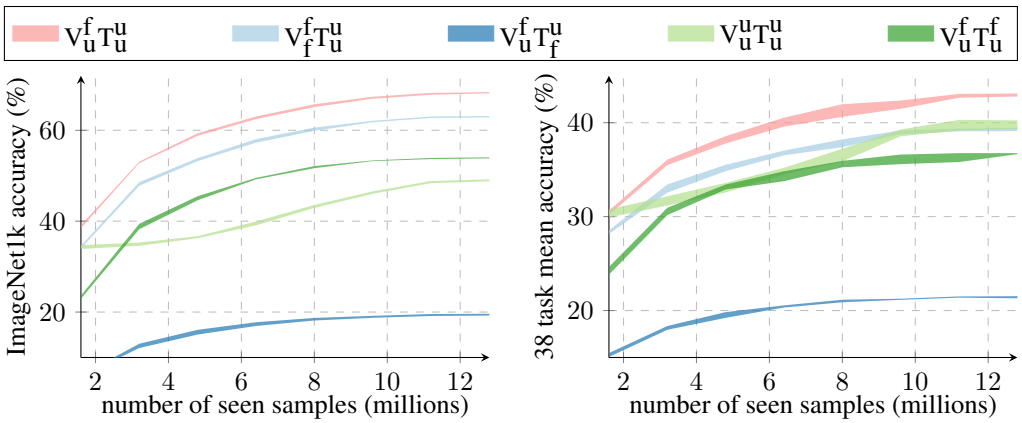

Figure 4: ImageNet1k (left) and 38 tasks (right) average accuracy over the number of samples seen. Each scenario is run with 5 seeds, line thickness is 2 standard deviations.

5 Conclusion

Determining which part of the pretrained models should be frozen and which parts should be unfrozen for fine-tuning is a nontrivial decision. For multimodal alignment of unimodal image and

text transformers, fine-tuning the CLS tokens is crucial. Compared to unfreezing everything, by freezing the vision transformer, unfreezing the CLS tokens, patch embedder, and the whole text model, performance on downstream evaluation tasks can be increased at a much lower training cost.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 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.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[5] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *URL https://arxiv. org/abs/2111.10050*, 2021.

[6] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[8] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[9] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

[10] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[13] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[14] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022.

[15] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[17] Common crawl. `https://commoncrawl.org`. Accessed on 26.09.2023.

[18] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[23] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[24] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[25] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.