Supplementary material for FFF: Fixing Flawed Foundations in contrastive pre-training results in very strong Vision-Language models

Adrian Bulat[1,2] Yassine Ouali[1] Georgios Tzimiropoulos[1,3]
[1]Samsung AI Center Cambridge, UK [2] Technical University of Iași, Romania
[3] Queen Mary University of London, UK

A. Additional comparisons with state-of-the-art

A.1. Zero-shot recognition on Open30M and Open70M datasets

To further showcase the scalability of our approach, we follow [2, 4], pretraining our method on a combination of 4 publicly available datasets, dubbed Open30M (see Tab. 6 for composition). The pretraining hyperparameters remain the same as for YFCC. Once trained, we evaluate it in a zero-shot manner on the same suite of 11 datasets. As the results from Tab. 1 show, our approach outperforms all prior methods, improving upon the prior best result of [2] by +4.7% aggregated over 11 datasets, including by +3.1% on ImageNet.

Finally, we extend the Open30M images dataset by adding RedCaps [1], OpenImages-8M [3] and YFCC-v1, creating Open70M. As the results from Tabs. 1 and 3 show, our approach scales well, with consistent gains for both zero-shot retrieval and classification.

A.2. Linear probe

In addition to zero-shot evaluation, we also present linear probe results in Tab. 2 for models pre-trained on YFCC15M and in Tab. 4 for models pre-trained on Open30M. Similar to zero-shot experiments, we use the `clip-benchmark` repository[1] to run these experiments. For each dataset, we cache the features of the training and test sets, and then use the training set's features and its ground-truth labels to train a linear layer on top. The linear linear is trained for 20 epochs using the standard cross-entropy loss and AdamW optimizer with a learning rate of 0.1, no weight decay, and a cosine learning rate scheduler. The trained linear layer is then used over the cached test features to obtain the accuracy. Similar to zero-shot experiments, our approach outperforms previous methods by large margins, *i.e*., +7.0% with YFCC15M pertaining (Tab. 2) and +6.2% with Open30M pertaining over 11 image classification datasets.

B. Additional ablation studies

Sensitivity to the threshold value: The selection of threshold values is intuitive, and the model is generally forgiving within a certain plateau of values. For S_{tt} and S_{ii}, they are simply set to high values to target nearly identical samples. For S_{it}, we start from the mean score of the positive pairs, which is 0.29, and explore a few adjacent values, noting that all values located in the same vicinity perform well as shown in Tab. 5.

C. Zero-shot classification prompts

For zero-shot recognition, we align with prior work [6, 7], using the same list of prompts. The full list is defined in Tab. 7.

D. Zero-shot retrieval evaluation considerations

As the synthetic captions are generated by models pre-trained on external data, a reasonable question to ask is wherever there is potential data leakage. For the Flick30k dataset, no such issues are present, as BLIP2 did not use any data from the training set of Flickr30k during any of its training phases. For MSCOCO, we note that only 100k out of 120M samples used for training BLIP2 were images from the COCO training set, hence the impact is likely minimal, if any. We note here that the current state-of-the-art method, ALIP, is subject to the same potential issue, as they also make use of synthetic captions produced by a model that was pre-trained on MSCOCO data (i.e. OFA).

[1] https://github.com/LAION-AI/CLIP_benchmark

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	ImageNet	Average
CLIP-ViT-B/32 [6]	Open30M	77.3	48.1	59.1	58.5	58.2	52.6	17.7	28.0	80.8	3.2	48.8	48.4
HiCLIP-ViT-B/32 [2]	Open30M	77.6	56.2	63.9	65.6	62.5	60.7	22.2	38.0	82.4	5.5	52.9	53.4
UniCLIP-ViT-B/32 [4]	Open30M	87.8	56.5	64.6	69.2	8.0	61.1	19.5	36.6	84.0	4.7	54.2	49.7
HiDeCLIP-ViT-B/32 [2]	Open30M	80.4	54.2	68.9	73.5	**66.1**	65.2	26.8	44.1	**87.8**	**7.2**	56.9	57.4
FFF-ViT-B/32 (Ours)	Open30M	**92.4**	**73.6**	**70.4**	**79.9**	64.1	**67.7**	**41.2**	**44.3**	84.1	5.2	**60.0**	**62.1**
FFF-ViT-B/32 (Ours)	Open70M	**92.7**	**73.7**	**79.8**	**78.8**	**68.3**	**68.7**	**47.3**	**51.1**	**86.5**	**5.3**	**65.9**	**65.3**

Table 1. Zero-shot classification performance on 11 downstream datasets. Results taken from [2].

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	Average
CLIP-ViT-B/32 [6]	YFCC15M	86.5	64.7	69.2	64.6	90.6	66.0	24.9	61.3	79.1	23.1	63.0
DeCLIP-ViT-B/32 [5]	YFCC15M	89.2	69.0	75.4	72.2	94.4	71.6	31.0	68.8	87.9	27.6	68.7
HiCLIP-ViT-B/32 [2]	YFCC15M	89.5	71.1	73.5	70.6	91.9	68.8	30.8	63.9	84.8	27.4	67.2
HiDeCLIP-ViT-B/32 [2]	YFCC15M	88.1	70.7	77.6	75.5	**95.6**	72.2	36.0	70.1	90.0	32.6	70.8
ALIP-ViT-B/32 [7]	YFCC15M	**94.3**	77.8	75.8	76.0	95.1	73.3	33.6	71.7	88.5	36.1	72.2
FFF-ViT-B/32 (Ours)	YFCC15M	93.9	**78.4**	**80.3**	**84.9**	94.7	**96.2**	**55.5**	**72.2**	**99.9**	**36.5**	**79.2**

Table 2. Linear probe classification performance on various downstream datasets. All models were pre-trained on YFCC15M. Results taken from [7].

		Text retrieval						Image retrieval					
		Flickr30k			MSCOCO			Flickr30k			MSCOCO		
Method	Pre-train dataset	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
FFF-ViT-B/32 (Ours)	YFCC-15M	**85.3**	**97.5**	**99.4**	**61.7**	**84.5**	**90.4**	**67.6**	**89.1**	**93.3**	**44.3**	**70.9**	**80.1**
FFF-ViT-B/32 (Ours)	Open30M	**87.9**	**99.2**	**99.6**	**64.2**	**85.8**	**91.7**	**72.0**	**91.4**	**94.9**	**46.4**	**72.6**	**81.6**
FFF-ViT-B/32 (Ours)	Open70M	**87.5**	**98.1**	**99.3**	**66.6**	**86.6**	**91.6**	**72.9**	**92.4**	**95.7**	**49.1**	**74.9**	**83.2**

Table 3. Zero-shot image-text retrieval on the test splits of Flickr30k and MSCOCO for models pretrained on YFCC-15M, Open30M and Open70M.

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	Average
CLIP-ViT-B/32 [6]	Open30M	92.0	74.7	78.8	80.7	93.7	72.6	55.9	71.4	88.6	29.7	73.8
HiCLIP-ViT-B/32 [2]	Open30M	92.8	75.8	80.5	81.3	94.4	73.6	59.4	72.2	90.3	33.6	75.4
DeCLIP-ViT-B/32 [5]	Open30M	93.1	76.9	82.0	82.7	96.0	74.9	59.8	74.5	92.6	32.7	76.5
HiDeCLIP-ViT-B/32 [2]	Open30M	92.7	75.6	82.9	83.3	95.7	75.6	62.8	74.5	92.0	35.8	77.1
FFF-ViT-B/32 (Ours)	Open30M	**96.6**	**84.1**	**83.8**	**87.4**	**95.7**	**97.3**	**74.1**	**75.5**	**99.9**	**38.7**	**83.3**

Table 4. Linear probe classification performance on various downstream datasets. All models were pre-trained on Open30M. Results taken from [2].

0.26	0.27	0.28	0.29	0.3
32.4	32.9	32.8	32.8	32.5

Table 5. **Effect of the S_{it} threshold (p_1):** Zero-shot evaluation on Imagenet in terms of Top-1 (%) accuracy.

Pre-train dataset	Number of examples
SBU	844,574
CC12M	10,503,723
CC3M	2,876,999
YFCC15M-V2	14,864,773
Open30M	29,090,069

Table 6. Number of examples per each training dataset. Open30M is the combination of all four datasets, *i.e.*, SBU, CC3M, CC12M and YFCC15M-V2.

CIFAR 10 & CIFAR 100			
a photo of a {label}.	a blurry photo of a {label}.	a black and white photo of a {label}.	a low contrast photo of a {label}.
a high contrast photo of a {label}.	a bad photo of a {label}.	a good photo of a {label}.	a photo of a small {label}.
a photo of a big {label}.	a photo of the {label}.	a blurry photo of the {label}.	a black and white photo of the {label}.
a low contrast photo of the {label}.	a high contrast photo of the {label}.	a bad photo of the {label}.	a good photo of the {label}.
a photo of the small {label}.	a photo of the big {label}.		
Food101			
a photo of {label}, a type of food.			
Caltech101			
a photo of a {label}.	a painting of a {label}.	a plastic {label}.	a sculpture of a {label}.
a sketch of a {label}.	a tattoo of a {label}.	a toy {label}.	a rendition of a {label}.
a embroidered {label}.	a cartoon {label}.	a {label} in a video game.	a plushie {label}.
a origami {label}.	art of a {label}.	graffiti of a {label}.	a drawing of a {label}.
a doodle of a {label}.	a photo of the {label}.	a painting of the {label}.	the plastic {label}.
a sculpture of the {label}.	a sketch of the {label}.	a tattoo of the {label}.	the toy {label}.
a rendition of the {label}.	the embroidered {label}.	the cartoon {label}.	the {label} in a video game.
the plushie {label}.	the origami {label}.	art of the {label}.	graffiti of the {label}.
a drawing of the {label}.	a doodle of the {label}.		
Stanford Cars			
a photo of a {label}.	a photo of the {label}.	a photo of my {label}.	i love my {label}!
a photo of my dirty {label}.	a photo of my clean {label}.	a photo of my new {label}.	a photo of my old {label}.
DTD			
a photo of a {label} texture.	a photo of a {label} pattern.	a photo of a {label} thing.	a photo of a {label} object.
a photo of the {label} texture.	a photo of the {label} pattern.	a photo of the {label} thing.	a photo of the {label} object.
FGVC Aircraft			
a photo of a {label}, a type of aircraft.	a photo of the {label}, a type of aircraft.		
Flowers102			
a photo of a {label}, a type of flower.			
Pets			
a photo of a {label}, a type of pet.			
SUN39			
a photo of a {label}.	a photo of the {label}.		
ImageNet			
a bad photo of a {label}.	a photo of many {label}.	a sculpture of a {label}.	a photo of the hard to see {label}.
a low resolution photo of the {label}.	a rendering of a {label}.	graffiti of a {label}.	a bad photo of the {label}.
a cropped photo of the {label}.	a tattoo of a {label}.	the embroidered {label}.	a photo of a hard to see {label}.
a bright photo of a {label}.	a photo of a clean {label}.	a photo of a dirty {label}.	a dark photo of the {label}.
a drawing of a {label}.	a photo of my {label}.	the plastic {label}.	a photo of the cool {label}.
a close-up photo of a {label}.	a black and white photo of the {label}.	a painting of the {label}.	a painting of a {label}.
a pixelated photo of the {label}.	a sculpture of the {label}.	a bright photo of the {label}.	a cropped photo of a {label}.
a plastic {label}.	a photo of the dirty {label}.	a jpeg corrupted photo of a {label}.	a blurry photo of the {label}.
a photo of the {label}.	a good photo of the {label}.	a rendering of the {label}.	a {label} in a video game.
a photo of one {label}.	a doodle of a {label}.	a close-up photo of the {label}.	a photo of a {label}.
the origami {label}.	the {label} in a video game.	a sketch of a {label}.	a doodle of the {label}.
a origami {label}.	a low resolution photo of a {label}.	the toy {label}.	a rendition of the {label}.
a photo of the clean {label}.	a photo of a large {label}.	a rendition of a {label}.	a photo of a nice {label}.
a photo of a weird {label}.	a blurry photo of a {label}.	a cartoon {label}.	art of a {label}.
a sketch of the {label}.	a embroidered {label}.	a pixelated photo of a {label}.	itap of the {label}.
a jpeg corrupted photo of the {label}.	a good photo of a {label}.	a plushie {label}.	a photo of the nice {label}.
a photo of the small {label}.	a photo of the weird {label}.	the cartoon {label}.	art of the {label}.
a drawing of the {label}.	a photo of the large {label}.	a black and white photo of a {label}.	the plushie {label}.
a dark photo of a {label}.	itap of a {label}.	graffiti of the {label}.	a toy {label}.
itap of my {label}.	a photo of a cool {label}.	a photo of a small {label}.	a tattoo of the {label}.

Table 7. The list of prompts used to evaluate the performance of zero-shot classification on 11 visual recognition datasets.

References

[1] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 1

[2] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pre-training with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 1, 2

[3] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 1

[4] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, 35:1008–1019, 2022. 1, 2

[5] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[6] 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. 1, 2

[7] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 1, 2