Revisiting Multimodal Representation in Contrastive Learning: From Patch and Token Embeddings to Finite Discrete Tokens (Supplementary Material)

Yuxiao Chen[1], Jianbo Yuan[2], Yu Tian[2], Shijie Geng[1,2], Xinyu Li[2],
Ding Zhou[2], Dimitris N. Metaxas[1], Hongxia Yang[3]
[1]Rutgers University [2]ByteDance Inc. [3]Zhejiang University

1. Pre-training Implementation Details

We implement the projecting function that maps patch or language token features to the FDT space as a fully-connected layer with GELU activation (see Section 3.2). Two different projecting functions are applied for mapping patch and language token features, respectively. We regularize the FDT using weight decay, with a rate of 0.1. We set the batch sizes as 4096, 8192, and 32768 when pretraining the models under the 15M, 30M, and 145M settings, respectively. To ensure a fair comparison with the DECLIP [12] and FILIP [21] models, we use the same data augmentation as these models when training the CLIP and CLIP+FDT models. Consequently, our reported results of the CLIP model on the 15M setting are better than those reported in the 15M benchmark [5]. We train ViT-B/32 based [7] models considering our limited computation resource. The input image resolution is 224×224, and the maximal input language token number is 77. Following [5], we apply the AdamW optimizer [15] with a weight decay rate of 0.1 during pre-training. The learning rate is first linearly increased to 0.001 with one epoch for warmup, and then decayed to 0 following the cosine strategy [14]. We use NVIDIA A100 GPUs for pre-training.

2. Downstream Implementation Details

2.1. Downstream Datasets

Image Classification Tasks. Following [12], we evaluate our method on 11 datasets, including CIFAR-10 [11], CIFAR-100 [11], SUN397 [20], Stanford Cars [10], FGVC Aircraft [16], Describable Textures [4], Oxford-IIIT Pets [18], Caltech-101 [9], Oxford Flowers 102 [17], Food-101 [3], and ImageNet-1K [6].
Image-Text Retrieval. Our method is tested on two standard benchmarks: Flickr30K [22] and MSCOCO [13]. For MSCOCO, we report the results on the 5K setting.
Non-Linear Probe task. We conduct the experiments on the VQAv2 dataset [2]. Following the standard protocol [8], we train the models with both training and validation data, and test the models on the test-dev set.

2.2. Implementation Details

Zero-shot Image Classification. For a fair comparison, we use the domain-specific prompts and category names proposed by CLIP [19]. Note that we do not report the results on the StanfordCars and Aircraft datasets, because the pertaining datasets contain few captions about the category names of these datasets. For example, only 0.04% and 0% of descriptions contain aircraft and car category names on the 15M setting.
Linear Probe Image Classification. We train a logistic regression classifier using L-BFGS, following CLIP [19]. We set the maximum iterations number to 1,000, and determine the L2 regularization weights following DECLIP's hyperparameter sweeping strategy [12]. We do not report the results on the ImageNet-1K dataset, due to the high computational cost of conducting hyperparameter sweeping on the dataset.
Non-linear Probe Task. The downstream task head consists of a fully-connected layer with GELU activation and a fully-connected layer. The extracted FDT features of images and questions are concatenated and then fed to the downstream task head to predict the answers. The encoders and FDT are frozen during the training. The downstream head is optimized by the AdamW optimizer [15]. We set the learning rate as 0.005, and decay it to 0 following the cosine strategy [14].

3. Completeness Probing Experiment Details

Given an image that contains N objects, its *matched sentence* is "An photo contains o_1, o_2 ..., o_{N-1}, and o_N", where o_i is the name of the i-th object in the images and all the objects are included. For the *partially matched sentence*, we randomly remove an object and use the remaining $N-1$ objects to construct a caption. For example, if the N-th object is removed, the partially matched sentence is "An photo contains o_1, o_2 ..., and o_{N-1}". We can construct N

partially matched sentences for the image, resulting in N *sentence pairs* for the image. In our experiments, we obtain the object presence information of images based on the object detection annotations of the MSCOCO [13] dataset. We construct 305,723 sentence pairs using all images in the MSCOCO training split.

4. FDT Visualization Details

We use the model pre-trained on the 145M setting for visualization because it achieves the best performance. To visualize an FDT token, we first calculate its relevance score between patches/language tokens following Equations 4 and 6 without using max-pooing. We then display the relevance scores between the FDT token and the images corresponding to the top-5 most relevant patches, since we find that the patches alone cannot fully convey the object information. We increase the resolution by reducing the patch stride to 4, following the method proposed in [1]. For text modality, we show the top-5 most relevant language tokens of the FDT token.

5. Additional Experiment Results

5.1. Text-to-Image Retrieval Cases

We further provide five cases for the text-to-image retrieval task in Figure 1. We have the same observation that the images retrieved by the CLIP+FDT well match the text queries, while those retrieved by the CLIP models often overlook important concepts mentioned in the text queries.

5.2. Visualization of Learned FDT

We present eight learned FDT in Figure 2. These cases further show that FDT can learn meaningful cross-modal correspondence.

5.3. Pretraining Data Scale

The results of the models pretrained with different scales of training data are reported in Table 1, 2, 3, and 4.

5.4. Image Encoder Architecture

To evaluate the influence of encoder architectures on our methods, we pre-trained the models with different image encoder architectures. The results for various downstream tasks are reported in in Table 5, 6, 7, and 8. We also report the computation costs when using different encoder architectures in Table 9.

5.5. FDT Number

The results of models trained with different FDT numbers are shown in Table 10, 11, 12, and 13.

5.6. Sparse Constraints

We report the results of the models trained with and without sparse constraint in Table 14, 15, 16, and 17.

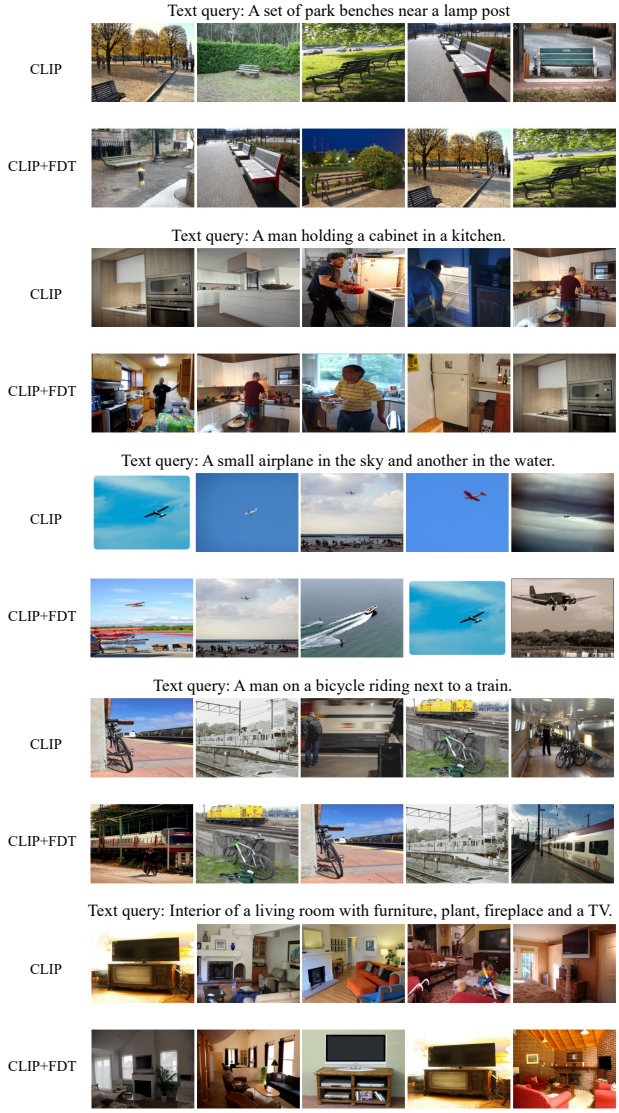

Figure 1. Examples show the top-5 retrieved images for the given text queries in the text-to-image retrieval task on MSCOCO.

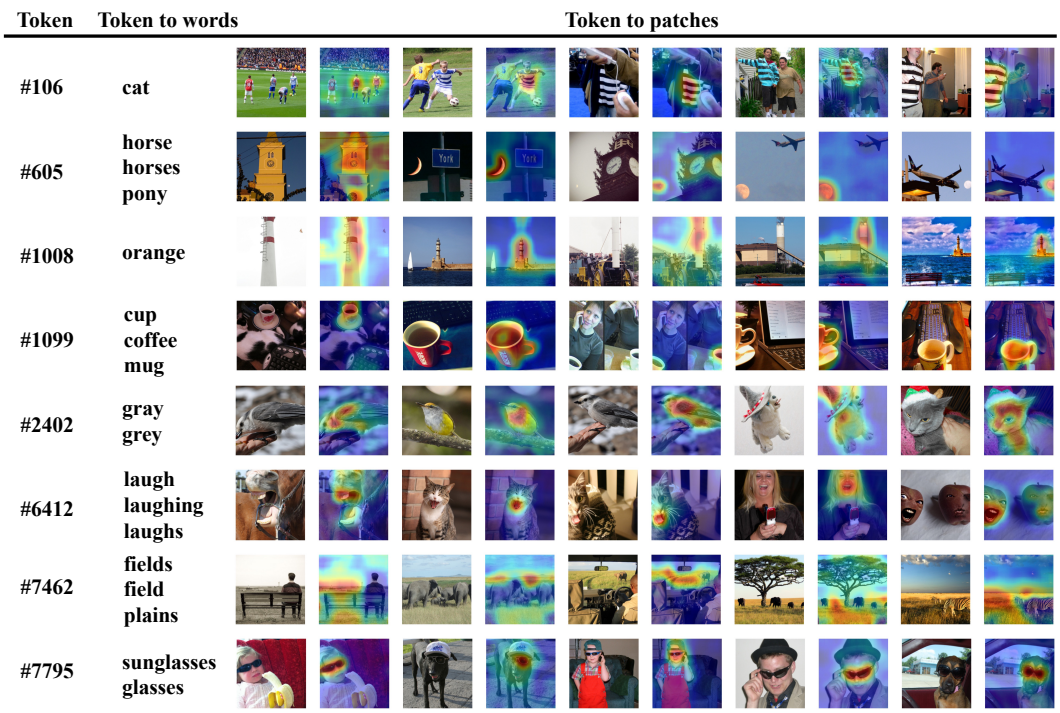

Figure 2. The top-5 most relevant image patches and text tokens of eight FDT tokens. Note that the redundant text tokens in the top-5 are removed. The color of the heatmap from blue to red denotes the relevance between patches and FDT from small to large.

	C10	C100	F101	PETS	FLOWE	SUN	DTD	CAL	IN	AVG
15M										
CLIP	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
CLIP+FDT	67.7	39.9	42.9	25.8	55.5	45.5	26.5	69.6	39.3	45.9 (↑ 4.6)
30M										
CLIP	77.2	48.1	59.1	58.4	58.2	52.6	28.0	80.8	48.8	56.8
CLIP+FDT	81.9	56.5	62.6	62.3	59.5	56.7	33.6	84.8	53.3	61.2(↑ 4.4)
145M										
CLIP	80.9	53.9	69.1	68.9	59.3	52.1	43.0	90.1	59.0	64.0
CLIP+FDT	87.1	63.7	73.5	77.0	65.0	56.2	47.7	90.5	60.4	69.0(↑ 5.0)

Table 1. Zero-shot image classification accuracy (%) when using different scales of training data. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	AIR	AVG
15M											
CLIP	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
CLIP+FDT	89.1	71.2	74.4	73.0	93.4	70.8	31.4	69.4	87.7	27.9	68.8 (↑ 1.3)
30M											
CLIP	92.0	74.7	78.8	80.7	93.7	72.6	55.9	71.4	88.6	29.7	73.8
CLIP+FDT	93.8	77.8	81.6	82.6	94.5	74.3	54.4	73.9	92.3	30.9	75.6 (↑ 1.8)
145M											
CLIP	95.2	80.6	86.1	87.5	96.5	76.3	87.6	77.2	94.7	39.5	82.1
CLIP+FDT	94.8	80.8	85.5	85.8	95.7	75.9	88.1	78.5	94.6	42.9	82.3 (↑ 0.2)

Table 2. Linear probing image classification accuracy (%) when using different scales of training data. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

	Flickr30K							MSCOCO						
	Image Retrieval			Text Retrieval				Image Retrieval			Text Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	rsum	R@1	R@5	R@10	R@1	R@5	R@10	rsum
15M setting														
CLIP	27.6	53.9	64.4	42.8	71.5	82.9	343.1	15.9	36.7	47.8	24.8	49.8	61.8	236.8
CLIP + FDT	32.6	58.6	68.5	51.0	78.3	87.5	376.5 (↑ 33.4)	19.4	40.8	51.9	29.6	55.3	66.1	263.1 (↑ 26.3)
30M setting														
CLIP	43.6	72.8	81.3	58.8	84.2	90.6	431.3	23.3	46.9	58.6	34.8	63.3	73.9	300.8
CLIP + FDT	52.5	78.7	86.4	70.8	90.8	95.0	474.2 (↑ 42.9)	28.3	53.3	64.3	43.0	69.0	79.2	337.1 (↑ 36.3)
145M setting														
CLIP	52.6	78.5	86.4	67.9	89.9	94.5	469.8	29.3	54.1	65.4	42.1	67.1	77.2	335.2
CLIP + FDT	56.3	80.7	87.6	75.9	93.6	95.3	489.4 (↑ 19.6)	31.0	55.7	66.7	46.4	71.9	81.3	353.0 (↑ 17.8)

Table 3. Zero-shot image-text retrieval results on the Flickr30K and MSCOCO (5K) datasets when using different scales of training data.

	y/n	number	other	overall
15M setting				
CLIP	67.7	31.9	33.6	47.5
CLIP + FDT	67.8	34.6	39.6	50.6 (↑ 3.1)
30M setting				
CLIP	69.7	34.8	37.8	50.6
CLIP + FDT	68.8	36.4	42.0	53.4 (↑ 2.8)
145M setting				
CLIP	70.9	36.5	41.7	53.1
CLIP + FDT	71.5	37.9	45.2	55.2 (↑ 2.1)

Table 4. Results of non-linear probing on VQA v2 dataset when using different scales of training data.

	C10	C100	F101	PETS	FLOW	SUN	DTD	CAL	IN	AVG
ViT-B/32	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
ViT-B/32+FDT	67.7	39.9	42.9	25.8	55.5	45.5	26.5	69.6	39.3	45.9 (↑ 4.6)
ViT-B/16	64.6	32.1	49.7	25.7	59.7	43.4	21.3	67.9	42.1	45.2
ViT-B/16+FDT	74.0	42.1	49.4	28.5	62.2	50.5	25.1	71.4	45.6	49.9 (↑ 4.7)
SwinV2-B	58.3	23.3	39.3	20.0	55.2	40.1	18.9	62.1	38.9	39.6
SwinV2-B+FDT	58.9	26.0	44.7	23.8	55.4	43.3	21.4	66.2	42.3	42.4 (↑ 2.8)

Table 5. Zero-shot image classification accuracy (%) when using different image encoder architectures. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	Air	AVG
ViT-B/32	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
ViT-B/32+FDT	89.1	71.2	74.4	73.0	93.4	70.8	31.4	69.4	87.7	27.9	68.8 (↑ 1.3)
ViT-B/16	89.2	69.5	80.3	75.1	95.9	73.4	33.4	71.5	88.3	32.0	68.8
ViT-B/16+FDT	89.3	71.6	82.3	75.8	96.1	74.2	34.0	71.8	88.6	29.3	71.3 (↑ 2.5)
SwinV2-B	85.6	65.1	78.5	71.4	94.3	72.3	30.8	69.4	85.9	32.1	68.5
SwinV2-B+FDT	86.8	67.5	80.5	75.6	94.8	73.1	33.4	72.7	88.9	34.0	70.7 (↑ 2.2)

Table 6. Linear probing image classification accuracy (%) when using different image encoder architectures. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

	Flickr30K							MSCOCO						
	Image Retrieval			Text Retrieval				Image Retrieval			Text Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	rsum	R@1	R@5	R@10	R@1	R@5	R@10	rsum
ViT-B/32	27.6	53.9	64.4	42.8	71.5	82.9	343.1	15.9	36.7	47.8	24.8	49.8	61.8	236.8
ViT-B/32+FDT	32.6	58.6	68.5	51.0	78.3	87.5	376.5 (↑ 33.4)	19.4	40.8	51.9	29.6	55.3	66.1	263.1 (↑ 26.3)
ViT-B/16	35.3	60.6	71.7	50.5	81.1	88.6	387.8	19.3	41.3	52.8	29.7	54.3	66.2	263.6
ViT-B/16+FDT	41.6	67.5	76.9	60.8	86.1	92.6	425.5(↑ 37.7)	23.4	46.7	58.0	35.3	60.4	71.6	295.4(↑ 31.8)
SwinV2-B	30.5	56.8	67.8	48.5	77.7	86.8	368.1	17.7	38.4	49.7	26.0	52.1	63.7	247.6
SwinV2-B+FDT	39.6	65.2	74.9	57.9	85.7	92.2	415.5(↑ 47.4)	22.3	44.9	56.2	33.8	60.1	71.0	288.3(↑ 40.7)

Table 7. Zero-shot image-text retrieval results on the Flickr30K and MSCOCO (5K) datasets when using different image encoder architectures.

	y/n	number	other	overall
ViT-B/32	67.7	31.9	33.6	47.5
ViT-B/32 + FDT	67.8	34.6	39.6	50.6 (↑ 3.1)
ViT-B/16	69.0	33.2	36.0	49.2
ViT-B/16 + FDT	72.0	37.6	42.9	54.3(↑ 5.1)
SwinV2-B	67.8	29.4	32.1	46.5
SwinV2-B + FDT	68.6	34.5	41.0	51.6(↑ 5.1)

Table 8. Results of non-linear probing on VQA v2 dataset when using different image encoder architectures.

	#param	FLOPs	Training time (s/iter)	Inference throughput (image-text pairs/s)
CLIP-ViT-B/32	151M	7.3G	0.50	808.5
CLIP-ViT-B/32+FDT	161M	9.4G	0.60	642.8
CLIP-ViT-B/16	150M	20.5G	1.15	315.7
CLIP-ViT-B/16+FDT	160M	25.1G	1.29	272.5
CLIP-Swin-B	151M	18.4G	1.41	258.3
CLIP-Swin-B+FDT	161M	20.5G	1.51	248.1

Table 9. Computation cost when using different image encoder architecture.

FDT size	C10	C100	F101	PETS	FLOW	SUN	DTD	CAL	IN	AVG
-	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
8192	**70.4**	**40.4**	38.3	19.9	51.3	42.8	16.6	68.1	37.8	42.8
16384	67.7	39.9	**42.9**	**25.8**	55.5	**45.5**	**26.5**	69.6	39.3	**45.9**
24576	69.0	39.1	41.9	24.2	**55.7**	44.4	21.8	**70.5**	**39.8**	45.2

Table 10. Zero-shot image classification accuracy (%) of models with different FDT sizes. The row whose FDT value is "-" represents the CLIP model. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

FDT size	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	Air	AVG
-	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
8192	89.1	70.3	72.8	70.7	**93.4**	70.1	29.6	68.5	87.2	27.5	67.9
16384	89.1	**71.2**	74.4	**73.0**	**93.4**	**70.8**	**31.4**	69.4	**87.7**	27.9	**68.8**
24576	**89.3**	71.0	**74.9**	71.2	**93.4**	70.6	30.1	**69.8**	87.2	**28.7**	68.6

Table 11. Linear probing image classification accuracy (%) of models with different FDT sizes. The row whose FDT value is "-" represents the CLIP model. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

	Flickr30K							MSCOCO						
	Image Retrieval			Text Retrieval				Image Retrieval			Text Retrieval			
FDT size	R@1	R@5	R@10	R@1	R@5	R@10	rsum	R@1	R@5	R@10	R@1	R@5	R@10	rsum
-	27.6	53.9	64.4	42.8	71.5	82.9	343.1	15.9	36.7	47.8	24.8	49.8	61.8	236.8
8192	32.7	58.3	68.7	50.6	77.4	86.9	374.6	18.5	40.4	51.7	29.1	53.6	64.8	258.1
16384	32.6	58.6	68.5	**51.0**	**78.3**	**87.5**	376.5	**19.4**	**40.8**	**51.9**	29.6	55.3	66.1	**263.1**
24576	**33.3**	**60.3**	**70.4**	50.4	78.1	86.0	**378.5**	18.6	40.3	51.8	**29.7**	**55.8**	**66.9**	**263.1**

Table 12. Zero-shot image-text retrieval results on the Flickr30K and MSCOCO (5K) datasets of models with different FDT sizes. The row whose FDT value is "-" represents the CLIP model.

FDT size	y/n	number	other	overall
-	67.7	31.9	33.6	47.5
8192	68.1	33.3	38.5	50.1
16384	67.8	34.6	39.6	50.6
24576	**68.7**	**35.2**	**40.3**	**51.4**

Table 13. Results of non-linear probing on VQA v2 dataset of models with different FDT sizes. The row whose FDT value is "-" represents the CLIP model.

	C10	C100	F101	PETS	FLOW	SUN	DTD	CAL	IN	AVG
CLIP	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
CLIP+FDT$_{\text{Softmax}}$ *	23.7	1.2	4.6	2.7	1.8	3.5	4.2	4.1	1.2	5.2
CLIP+FDT$_{\text{Sparsemax}}$ *	59.9	24.7	17.3	20.9	35.1	31.2	20.8	56.8	25.0	32.4
CLIP+FDT$_{\text{Softmax}}$	68.7	36.9	35.5	27.9	53.8	43.8	23.1	66.6	38.6	43.9
CLIP+FDT$_{\text{Sparsemax}}$	67.7	39.9	42.9	25.8	55.5	45.5	26.5	69.6	39.3	45.6

Table 14. Zero-shot image classification accuracy (%) of models trained with (Sparsemax) and without (Softmax) sparse constraints. The rows marked with "*" are the results when using FDT weights as features. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	Air	AVG
CLIP	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
CLIP+FDT$_{\text{Softmax}}$	88.0	71.7	74.8	71.9	93.8	70.4	30.5	69.8	87.3	28.6	68.7
CLIP+FDT$_{\text{Sparsemax}}$	89.1	71.2	74.4	73.0	93.4	70.8	31.4	69.4	87.7	27.9	68.8

Table 15. Linear probing image classification accuracy (%) of models trained with (Sparsemax) and without (Softmax) sparse constraints. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

	Flickr30K							MSCOCO						
	Image Retrieval			Text Retrieval				Image Retrieval			Text Retrieval			
FDT size	R@1	R@5	R@10	R@1	R@5	R@10	rsum	R@1	R@5	R@10	R@1	R@5	R@10	rsum
CLIP	27.6	53.9	64.4	42.8	71.5	82.9	343.1	15.9	36.7	47.8	24.8	49.8	61.8	236.8
CLIP+FDT$_{\text{Softmax}}$ *	5.4	12.0	16.3	1.7	3.8	6.3	45.5	2.4	6.8	9.7	0.8	2.4	4.1	26.2
CLIP+FDT$_{\text{Sparsemax}}$ *	10.5	29.8	39.2	32.5	59.8	70.6	242.4	6.0	16.5	24.1	18.3	40.5	52.1	157.5
CLIP+FDT$_{\text{Softmax}}$	33.3	60.7	69.5	47.9	78.0	88.2	377.6	19.2	40.3	51.7	28.3	53.8	65.5	258.8
CLIP+FDT$_{\text{Sparsemax}}$	32.6	58.6	68.5	51.0	78.3	87.5	376.5	19.4	40.8	51.9	29.6	55.3	66.1	263.1

Table 16. Zero-shot image-text retrieval results on the Flickr30K and MSCOCO (5K) datasets of models trained with (Sparsemax) and without (Softmax) sparse constraints. The rows marked with "*" are the results when using FDT weights as features.

	y/n	number	other	overall
CLIP	67.7	31.9	33.6	47.5
CLIP+FDT$_{\text{Softmax}}$	65.7	31.9	36.2	47.9
CLIP+FDT$_{\text{Sparsemax}}$	67.8	34.6	39.6	50.6

Table 17. Results of non-linear probing on VQAv2 dataset of models trained with (Sparsemax) and without (Softmax) sparse constraints.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[5] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022.

[6] 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.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] 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.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[16] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[18] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[19] 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.

[20] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[21] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[22] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.