Supplement Material of Paper "Learning Visual Representation from Modality-Shared Contrastive Language-Image Pre-training"

Haoxuan You[1⋆], Luowei Zhou[2⋆], Bin Xiao[2⋆], Noel Codella[2⋆], Yu Cheng[2], Ruochen Xu[2], Shih-Fu Chang[1], and Lu Yuan[2]

[1] Columbia University, New York, USA
[2] Microsoft Cloud and AI, Redmond, USA
{hy2612,sc250}@columbia.edu
{luozhou,bixi,ncodella,yu.cheng,ruox,luyuan}@microsoft.com

Table 1: Setting of Early Specialization when ViT-B/16 as visual backbone, N*N means 2D kernel size of CNNs.

Module	Stride	Dim	Resolution
3*3 Conv	2	2→48	224→112
Residual 3*3 Conv	2	48→96	112→56
Residual 3*3 Conv	2	96→192	56→28
Residual 3*3 Conv	2	192→384	28→14
Residual 3*3 Conv	1	384→768	14→14
1*1 Conv	1	768→768	14→14
Total # Parameters		4.5M	

Table 2: Setting of Early Specialization when ViT-B/16 as visual backbone, N*N means 2D kernel size of CNNs.

Parallel Module	Adapter Module	Fusion Layer	Resolution
3*3 Conv	8*8 DWConv	2	224→112
Bottleneck 3*3 Conv	4*4 DWConv	4	112→56
Bottleneck 3*3 Conv	2*2 DWConv	6	56→28
Bottleneck 3*3 Conv	1*1 DWConv	8	28→14
Bottleneck 3*3 Conv	1*1 DWConv	10	14→14
Total # Parameters		3.9M	

⋆ Equal Contribution

Table 3: Linear probing results on 24 datasets.

Datasets	CLIP (ViT-B32)	MS-CLIP-S (B32)	Δ
Food-101	68.5	**76.4**	+ 4.7
SUN397	62.0	**67.8**	+ 5.8
Stanford Cars	70.7	**79.1**	+ 8.4
FGVC Aircraft	38.6	**45.4**	+ 6.8
Pascal Voc 2007	80.1	**83.9**	+ 3.8
Describable Texture (dtd)	67.9	**75.1**	+ 7.2
Oxford-IIIT Pets	69.4	**77.4**	+ 8.0
Caltech-101	86.2	**88.9**	+ 2.7
Oxford Flowers 102	89.2	**93.5**	+ 4.3
MNIST	97.1	**98.1**	+ 1.0
Facial Emotion Recognition	56.8	**57.2**	+ 0.4
STL-10	93.8	**95**	+ 1.2
GTSRB	**86.4**	83.5	− 2.9
PatchCamelyon	81.0	**81.1**	+ 0.1
UCF101	70.8	**74.7**	+ 3.9
CIFAR-10	**93.5**	92.0	− 1.5
CIFAR-100	**78.0**	74.9	− 3.1
Hateful Memes	50.6	**52.0**	+ 1.4
ImageNet	59.1	**66.5**	+ 7.4
Country211	13.8	**16.4**	+ 2.6
EuroSAT	**95.1**	94.7	− 0.4
Kitti-distance	**44.4**	37.6	− 6.8
Rendered-SST2	56.8	**59.7**	+ 2.9
Resisc45	83.0	**87.5**	+ 4.5
Avg.	70.5	**73.3**	+ 2.8

1 Modality-Specific Auxiliary Module Configuration

When visual backbone is ViT-B/16, we slight adjust the convolution kernels and strides in Early Specialization and Efficient Parallel Branch. The detailed configuration of those two are shown in Tab. 1 and Tab. 2.

2 Detailed Linear Probing Results When Pre-trained on Laion-20M

The results of linear probing on 24 various datasets with models pre-trained on Laion-20M are shown in Tab. 3. Our MS-CLIP-S can outperform vanilla CLIP on 19 datasets with an average improvement of 2.7%.

3 Zero-shot Evaluation on 24 datasets

We further conduct zero-shot evaluation on all 24 datasets following the same configuration in CLIP. The complete result is shown in Tab. 4. Our MS-CLIP-S

Table 4: Zero-shot Eval. of models pre-trained on YFCC-22M and LAION-20M. B32 denotes using ViT-B/32 as visual backbone and B16 denotes using ViT-B/16 as visual backbone.

Datasets	YFCC-22M						LAION-20M		
	CLIP (B32)	MS-CLIP-S (B32)	Δ	CLIP (B16)	MS-CLIP-S (B16)	Δ	CLIP (B32)	MS-CLIP-S (B32)	Δ
Food-101	34.4	**41.1**	+6.7	39.8	**40.7**	+0.9	47.1	**56.3**	+9.2
SUN397	40.4	**42.1**	+1.7	37.6	**42.7**	+5.0	40.2	**47.5**	+7.3
Stanford Cars	1.3	**1.5**	+0.2	1.0	**1.9**	+0.9	13.6	**16.5**	+2.9
FGVC Aircraft	2.1	**2.3**	+0.3	**2.7**	2.5	-0.2	3.1	**4.1**	+1
Pascal Voc 2007	44.6	**48.1**	+3.5	45.1	**48.6**	+3.5	43.8	**48.6**	+4.8
Describable Texture (dtd)	13.4	**14.6**	+1.3	14.4	**19.5**	+5.1	26.7	**31.4**	+4.7
Oxford-IIIT Pets	**11.9**	8.7	-3.2	11.2	**11.3**	+0.1	50.6	**61.4**	+1.0
Caltech-101	**21.7**	19.3	-2.4	21.1	**22.9**	+1.8	27.2	**28.7**	+1.5
Oxford Flowers 102	35.4	**40.6**	+5.1	38.5	**40.8**	+2.3	33	**36.5**	+3.5
MNIST	9.9	**10.0**	+0.1	9.7	**10.4**	+0.7	17.6	**25.6**	+8
Facial Emotion Recognition	16.8	**19.8**	+3.0	**17.1**	12.4	-4.6	19.6	**23.4**	+3.8
STL-10	**89.9**	87.4	-2.5	86.8	**91.8**	+5.0	88.4	**90**	+1.6
GTSRB	7.6	**9.0**	+1.4	4.8	**11.8**	+7.0	**22.6**	15.3	-7.3
PatchCamelyon	**50.9**	50.0	-0.9	48.0	**53.9**	+5.9	**52.3**	50.4	-1.9
UCF101	**32.4**	30.4	-2.1	33.5	**34.4**	+0.9	39	**41.8**	+2.8
CIFAR-10	**79.4**	70.2	-9.1	**80.2**	73.0	-7.2	**85.1**	81.7	-3.4
CIFAR-100	4.6	**4.8**	+0.2	**4.3**	3.1	-1.2	**6.9**	5.2	-1.7
Hateful Memes	**49.6**	48.7	-0.9	49.7	**52.8**	+3.1	**53.5**	50.8	-2.7
ImageNet	32.2	**36.7**	+4.5	36.9	**39**	+4.7	35.5	**40.2**	+4.7
Country211	1.7	**2.2**	+0.4	2.0	**2.1**	+0.1	5.6	**7**	+1.4
EuroSAT	**16.7**	6.6	-10.1	6.1	**14.8**	+8.7	5.6	**5.8**	+0.2
Kitti-distance	13.2	**33.9**	+20.7	19.3	**38.0**	+18.7	**31.6**	27.8	-3.8
Rendered-SST2	**51.7**	49.9	-1.8	49.9	**50.2**	+0.3	47.9	**50.5**	+2.6
Resisc45	**24.4**	21.2	-3.2	**29.8**	28.4	-1.5	35.3	**37.7**	+2.4
# Win	10	**14**	+4	5	**19**	+14	6	**18**	+12
Avg.	28.5	**29.1**	+0.6	28.7	**31.1**	+2.4	34.6	**36.8**	+2.2

consistently outperforms CLIP in different pre-training datasets and backbone models. When pre-trained on LAION-20M, our MS-CLIP-S outperforms CLIP on 18 out of 24 datasets with an average gain of 2.2%. When pre-trained on YFCC-22M with ViT-B/16 as backbone, the average gain is 2.4% with outperforming on 19 out of 24 datasets. However, when pre-trained on YFCC-22M with ViT-B/32 as backbone, the overall improvement is not that significant. We hypothesize that because of a weaker baseline, the performances in many datasets are very low and the numerical fluctuation influence a lot.

4 More Ablations

4.1 Ablation on Sharing Attention and FFN individually

We further conduct experiments where either FFN or Attn is shared while others are modality-specific. As in Tab. 5, we found that still sharing both gives better result than individual sharing. We infer that it's probably because the attention modules' output is input into FFN modules, which makes them strongly coupled.

Table 5: Experimental results of sharing Attn. and FFN individually in Transformer layer. LN1 denotes the LN before Attn. LN2 denotes the LN before FFN.

Text Width	# Params	Shared Module	Non-Shared Module	IN Zero-shot Acc(%)
768	126M	Attn, FFN	LN1, LN2	**32.99**
768	154M	FFN	Attn, LN1, LN2	30.40
768	182M	Attn	FFN, LN1, LN2	26.12

Table 6: Ablation on whether using DWConv in adapters.

Model	# Params	IN Zero-shot Acc(%)
MS-CLIP-S	132M	**36.66**
··· w/o DWConv	131M	33.94

4.2 Ablation on Depth-Wise Conv in adapters

The Depth-Wise Conv (DWConv) can gather spatial context features with 2D kernels and resize image feature map, while FFN/BottleneckFFN is applied point-wise without context. To verify the importance of spatial context, we replace DWConv with average pooling + FFN (average pooling's kernel size, stride, padding are same as DWConv) which performs worse than DWConv by 2.7% in IN ZS accuracy, as shown in Tab. 6.