iCLIP: Bridging Image Classification and Contrastive Language-Image Pre-training for Visual Recognition

—— Supplementary Material ——

Yixuan Wei[1,2], Yue Cao[2], Zheng Zhang[2], Houwen Peng[2], Zhuliang Yao[1,2], Zhenda Xie[1,2]
Han Hu[2], Baining Guo[2]
[1]Tsinghua University [2]Microsoft Research Asia

Table 1. Ablation study on combination ratio of two kinds of datasets with the same number of samples. Models are pre-trained on YFCC-14M (YFCC) and ImageNet21K (IN-21k), and evaluated on zero-shot classification on IN-1K and zero-shot cross-modal retrieval on MS-COCO. We sample 50% images from both datasets in our mainscript as default.

#	Training Data	Method	IN-1K	COCO-IR	COCO-TR
1	100% YFCC + 0% IN-21k	CLIP	30.1	12.5	21.2
2	90% YFCC + 10% IN-21k	iCLIP	40.9	13.9	25.5
3	75% YFCC + 25% IN-21k	iCLIP	43.9	15.2	**27.5**
4	50% YFCC + 50% IN-21k	iCLIP	**45.9**	**15.5**	27.2
5	25% YFCC + 75% IN-21k	iCLIP	44.8	14.1	27.1
6	10% YFCC + 90% IN-21k	iCLIP	44.2	14.9	25.9

A. Dataset size ratio between classification and contrastive learning.

We conducted an ablation study for the effect of dataset size ratio on YFCC-14M and IN-21k datasets. The results, shown in the Tab. 1, indicate that our framework performs well with a broad range of data ratio configurations (10%-90%). The best performance is achieved when the sampling ratio is 50%:50%, indicating a sweet spot.

B. Comparison with UniCL [11] on multi-modal retrieval

In Tab. 3 of the main manuscript, we have compared iCLIP with UniCL on IN-1K and 14 datasets zero-shot classification. Here, we include results on zero-shot multi-modal retrieval in Tab. 2, using Flickr30K [12] (1K test set) and MSCOCO [6] (5K test set). Our method performs also better than UniCL on cross-modal retrieval benchmarks, since that the dictionary enhancement class names close the label granularity gap between the original class names (one or few words) and the alt-texts (complete sentences).

C. Setups for fine-tuning on down-stream tasks

For semantic segmentation, we conduct the experiment on ADE20K [13] dataset and report single scale mIoU on validation set. We utilize MaskFormer [1] as our base framework and adopt its default training recipe except for setting window size to 7. For object detection, we fine-tune the models on LVIS v1 [2] with Faster R-CNN [10], following the settings in Swin [7]. LVIS includes 1203 categories with an unbalanced distribution. We report single scale validation mAP^{box} on all categories and rare categories, respectively, under 2x schedule (24 epochs) with multi-scale training (shorter size between 480 and 800). We also evaluate on the video action recognition task on Kinetics-400 (K400) [4] dataset for 30 epochs, following the same recipe in Video Swin Transformer [8]. Top-1 accuracy is reported.

D. Detailed results on zero-shot classification

We compare iCLIP with CLIP [9] and OpenCLIP [3] on Kornblith 12-dataset benchmark [5] in the main body. Table 3 presents the detailed results on each dataset.

References

[1] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1

[3] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 1, 2

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola,

Table 2. Comparison with UniCL. Models are pre-trained from scratched with 32 epochs, following UniCL [11]. ‡ denotes for our reproduction. COCO and Flickr stands for MSCOCO [6] and Flickr30K [12]. IR and TR stands for image retrieval and text retrieval, and top-1 recall is reported. Models with the datasets of YFCC-14M and IN-22K are excluded, because the UniCL model is not publicly available.

#	Training Data	Method	Flickr30K-IR	Flickr30K-TR	MSCOCO-IR	MSCOCO-TR
1	YFCC-14M + IN-21K (half)	UniCL [11]	21.5‡	37.9‡	12.5‡	21.2‡
2	YFCC-14M + IN-21K (half)	iCLIP	**31.9**	**49.8**	**15.5**	**27.2**
3	YFCC-14M + IN-21K	UniCL [11]	34.0‡	50.3‡	17.7‡	28.0‡
4	YFCC-14M + IN-21K	iCLIP	**37.1**	**55.7**	**18.5**	**30.7**

Table 3. Detailed comparisons of zero-shot classification with CLIP and OpenCLIP on Kornblith 12-dataset classification benchmark [5].

Methods	*Food101*	*CIFAR10*	*CIFAR100*	*Birdsnap*	*SUN397*	*Stanford Cars*	*FGVC Aircraft*	*VOC2007*	*DTD*	*Oxford Pets*	*Caltech101*	*Flowers102*	***Average***
CLIP-ViT-B/16 [9]	89.2	91.6	68.7	39.1	65.2	65.6	27.1	83.9	46.0	88.9	89.3	70.4	68.8
OpenCLIP-ViT-B/16 [3]	86.1	91.7	71.4	50.2	69.4	83.7	17.7	82.9	50.8	89.3	91.7	66.6	70.9
iCLIP	82.7	94.8	78.4	48.5	62.9	63.1	8.4	84.5	62.9	87.9	92.1	81.3	70.6

Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666, 2019. 1, 2

[6] 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. 1, 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 10012–10022, 2021. 1

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 1

[9] 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

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[11] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19163–19173, June 2022. 1, 2

[12] 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. 1, 2

[13] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 1