RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-training Supplementary Material

Chen-Wei Xie*, Siyang Sun*, Xiong Xiong*, Yun Zheng, Deli Zhao, Jingren Zhou
Alibaba Group
{eniac.xcw,siyang.ssy,moxiong.xx,zhengyun.zy}@alibaba-inc.com
zhaodeli@gmail.com, jingren.zhou@alibaba-inc.com

Appendix

This appendix is organized as follows.

- In Section A, we provide details of the downstream datasets we used.
- In Section B, we list the prompt templates used for zero-shot classification experiments.
- In Section C, we describe the training details of our linear probe experiments.
- In Section D, we provide more implementation details of the $MultiheadAttn$ used in the proposed Retrieval Augmented Module (RAM).
- In Section E, we provide more visualization to show that RAM is robust to the quality of the retrieved image-text pairs.

A. Downstream Datasets

We have 12 widely used downstream datasets: ImageNet [4], ImageNet V2 [13], CIFAR 10 [8], CIFAR 100 [8], Caltech 101 [5], Oxford Pets [11], SUN 397 [14], Food 101 [1], DTD [3], Stanford Dogs [7], COCO [2] and LVIS [6]. Table 1 summarizes the details of these datasets. For ImageNet V2, we use the same training data of ImageNet for the linear probe classification experiments. For COCO and LVIS, we only use them to evaluate the zero-shot ROI classification, thus we don't need training data. The classification datasets use classification accuracy as evaluation metric, except for Caltech 101 and Oxford Pets, which use averaged per-class accuracy. The detection datasets use average precision as evaluation metric.

Table 1. Details of downstream datasets.

Dataset	#Classes	#Train	#Test	Metric
ImageNet	1,000	1,281,167	50,000	accuracy
ImageNet V2	1,000	–	50,000	accuracy
CIFAR 10	10	50,000	10,000	accuracy
CIFAR 100	100	50,000	10,000	accuracy
Caltech 101	102	3,060	6,085	mean-per-class
Oxford Pets	37	3,680	3,669	mean-per-class
SUN 397	397	19,850	19,850	accuracy
Food 101	102	75,750	25,250	accuracy
DTD	47	3,760	1,880	accuracy
Stanford Dogs	120	12,000	8,580	accuracy
COCO	81	–	5,000	average precision
LVIS	1,203	–	5,000	average precision

B. Prompt Engineering

Following previous works [9, 10, 12], we extend the category names into sentences with prompts such as "a photo of {label}." before feeding them into the text encoders. For a fair comparison, we adopt the same prompts used in CLIP [12]. Specifically, for Oxford Pets, we use "a photo of a {label}, a type of pet.", while for Food 101 dataset, we use 'a photo of a {label}, a type of food.". For the other datasets, we use 80 prompt templates as shown in Figure 1. For a given category name, we average the embeddings of different prompted sentences, and conduct L2 normalization to obtain the final category embedding.

C. Training Details of Linear Probe

We freeze the pre-trained image encoder and append a linear classifier after it for linear probe classification. During training, we apply data augmentation to the input image. Concretely, we random crop a 224×224 patch from input image, then conduct random horizontal flip. During testing, we resize the shorter size to 224 then center crop a 224×224 patch as input image. We train the classifier for 90 epochs except for the ImageNet dataset, for which we train 10 epochs in total due to the large data volume. The learning rate follows a cosine decay schedule with initial learning rate equal to 0.1. We use SGD with momentum for

*indicates equal contribution.

a bad photo of a {label}.	a close-up photo of a {label}.	the origami {label}.	a jpeg corrupted photo of the {label}.
a photo of many {label}.	a black and white photo of the {label}.	the {label} in a video game.	a good photo of a {label}.
a sculpture of a {label}.	a painting of the {label}.	a sketch of a {label}.	a plushie {label}.
a photo of the hard to see {label}.	a painting of a {label}.	a doodle of the {label}.	a photo of the nice {label}.
a low resolution photo of the {label}.	a pixelated photo of the {label}.	a origami {label}.	a photo of the small {label}.
a rendering of a {label}.	a sculpture of the {label}.	a low resolution photo of a {label}.	a photo of the weird {label}.
graffiti of a {label}.	a bright photo of the {label}.	the toy {label}.	the cartoon {label}.
a bad photo of the {label}.	a cropped photo of a {label}.	a rendition of the {label}.	art of the {label}.
a cropped photo of the {label}.	a plastic {label}.	a photo of the clean {label}.	a drawing of the {label}.
a tattoo of a {label}.	a photo of the dirty {label}.	a photo of a large {label}.	a photo of the large {label}.
the embroidered {label}.	a jpeg corrupted photo of a {label}.	a rendition of a {label}.	a black and white photo of a {label}.
a photo of a hard to see {label}.	a blurry photo of the {label}.	a photo of a nice {label}.	the plushie {label}.
a bright photo of a {label}.	a photo of the {label}.	a photo of a weird {label}.	a dark photo of a {label}.
a photo of a clean {label}.	a good photo of the {label}.	a blurry photo of a {label}.	itap of a {label}.
a photo of a dirty {label}.	a rendering of the {label}.	a cartoon {label}.	graffiti of the {label}.
a dark photo of the {label}.	a {label} in a video game.	art of a {label}.	a toy {label}.
a drawing of a {label}.	a photo of one {label}.	a sketch of the {label}.	itap of my {label}.
a photo of my {label}.	a doodle of a {label}.	a embroidered {label}.	a photo of a cool {label}.
the plastic {label}.	a close-up photo of the {label}.	a pixelated photo of a {label}.	a photo of a small {label}.
a photo of the cool {label}.	a photo of a {label}.	itap of the {label}.	a tattoo of the {label}.

Figure 1. The prompt templates used for zero-shot classification.

optimization. Weight decay is not used in our experiments. The batch size is set to 128.

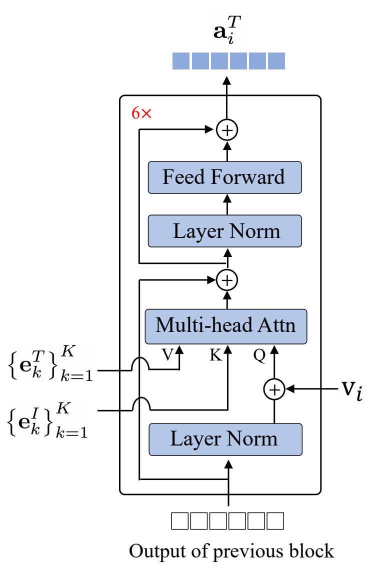

Figure 2. Implementation details of the Equation 2.

D. More Implementation Details of RAM

Equation 2 and Equation 3 in Section 3.2 of the main paper adopt $MultiheadAttn$ blocks to aggregate reference embeddings $\{\mathbf{e}_k^T\}_{k=1}^K$ and $\{\mathbf{e}_k^I\}_{k=1}^K$ for input $\mathbf{v}_i$. In this section, we take Equation 2 for example and provide more implementation details of it. As shown in Figure 2, the $MultiheadAttn$ block used in Equation 2 contains not only multi-head attention layer, but also layer normalization, feed forward block and short-cut connections. Given $\mathbf{v}_i$, RAM scans all reference image embeddings $\{\mathbf{e}_k^I\}_{k=1}^K$ and gathers related textual information from $\{\mathbf{e}_k^T\}_{k=1}^K$ into a new embedding $\mathbf{a}_i^T$. This process can be repeated for several times to iteratively refine $\mathbf{a}_i^T$, each block takes $\mathbf{v}_i$, $\{\mathbf{e}_k^T\}_{k=1}^K$, $\{\mathbf{e}_k^I\}_{k=1}^K$ and previous block's output $\mathbf{a}_i^T$ as inputs, then update $\mathbf{a}_i^T$ as output. The initial $\mathbf{a}_i^T$ fed into the first block is set to all-zero embedding.

E. More Visualization

In this section, we provide more visualization results of reference retrieval described in Section 3.2. As shown in Figure 3, the retrieval process may not return image-text pairs that provide description about the ground-truth category. However, the proposed RAM is robust to the quality of the retrieval results and still produces correct predictions. Specifically, for the first retrieval example in Figure 3, the input image is a photo of a *taxicab*, although the retrieved images are similar to the input image, their corresponding texts are not descriptions about *taxicab*. Since the proposed RAM not only depends on the retrieved image-text pairs, but also takes the embedding of input image into consideration, we can still give correct prediction for this case.

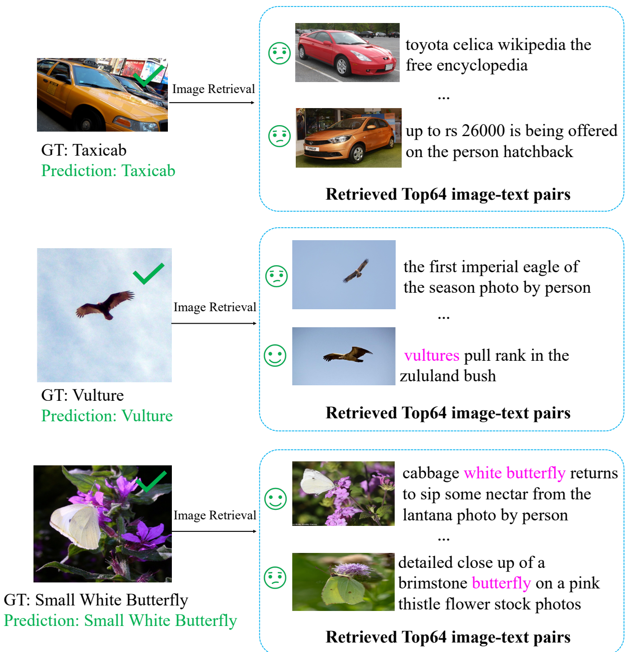

Figure 3. More visualization results of reference retrieval.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Eur. Conf. Comput. Vis.*, pages 446–461, 2014. 1

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, 2015. 1

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3606–3613, 2014. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 1

[5] 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 *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 178–178, 2004. 1

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5356–5364, 2019. 1

[7] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011. 1

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[9] 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. In *Int. Conf. Learn. Represent.*, 2022. 1

[10] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In *Eur. Conf. Comput. Vis.*, volume 13686, pages 529–544, 2022. 1

[11] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3498–3505, 2012. 1

[12] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 1

[13] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Int. Conf. Mach. Learn.*, volume 97, pages 5389–5400, 2019. 1

[14] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. SUN database: Exploring a large collection of scene categories. *Int. J. Comput. Vis.*, 119(1):3–22, 2016. 1