MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining

Xiaoyi Dong[1*†], Jianmin Bao[2*], Yinglin Zheng[3], Ting Zhang[2], Dongdong Chen[4,†], Hao Yang[2], Ming Zeng[3], Weiming Zhang[1], Lu Yuan[4], Dong Chen[2], Fang Wen[2], Nenghai Yu[1]
[1]University of Science and Technology of China [2]Microsoft Research Asia
[3]Xiamen University [4]Microsoft Cloud + AI
{dlight@mail., zhangwm@, ynh@}.ustc.edu.cn cddlyf@gmail.com
{jianbao, ting.zhang, luyuan, doch, fangwen}@microsoft.com
{zhengyinglin@stu., zengming@}xmu.edu.cn yanghao.alexis@foxmail.com

A. More Experiment

Comparison over small model and small dataset. As some baselines report ViT-B/32 instead of ViT-B/16, in order to compare, we further experiment MaskCLIP with a smaller model ViT-B/32 and report the zero-shot performance on ImageNet-1K. As shown in Table 1 left, our MaskCLIP outperforms the combination [5] of two recent strong methods DeCLIP [9] and FILIP [17]. We also investigate the performance on a smaller dataset CC3M [13] (we use ViT-B/16 here in coherency with previous experiments). Table 1(right part) shows that MaskCLIP achieves consistent gain.

ViT-B/32	0-Shot	Lin	FT	CC3M	0-Shot	Lin	FT
CLIP	26.1	60.5	74.3	CLIP	17.1	53.3	78.5
DeFILIP	36.4	—	—	SLIP	23.0	65.4	81.4
MaskCLIP	38.5	69.1	79.2	MaskCLIP	24.4	66.1	82.5

Table 1. Results of zero-shot performance on ImageNet-1K when pretrained with ViT-B/32 model(left) or CC3M dataset(right) .

Ablation on distillation loss. Here we further study the effectiveness of each component in the distillation loss. We start from CLIP+MAE and add three components of the distillation loss one by one. We find that 1) using the feature as the prediction target improves all metrics; 2) using EMA model gets better performance; 3) the MLM loss improves all the vision-language tasks.

	0-Shot	FT	Seg	Det	I2T/T2I
CLIP+MAE (baseline)	42.1	83.2	49.1	44.5/40.4	57.3/41.1
+ Feature prediction	42.6	83.4	49.9	45.1/40.6	62.3/41.4
+ EMA model	42.8	83.6	50.4	45.5/40.9	65.0/41.6
+ MLM loss	44.5	83.6	50.5	45.4/40.9	70.1/45.6

Table 2. Component ablation of the distillation loss.

*Equal contribution, † Corresponding Author
†Work done during an internship at Microsoft Research Asia

B. Experiment detail

Pre-training We train our proposed MaskCLIP from scratch and training for 25 epochs, the batch size is fixed to 4096 for all the experiments. We use 32 V100 for training with 128 samples per GPU. We use the AdamW [10] optimizer with weight decay 0.1. The learning rate is set to $1e^{-3}$ with one epoch warm-up and decay to $1e^{-5}$ followed by a cosine schedule. The masks used in the mask self-distillation branch are random mask with a mask ratio of 75%. The EMA weight is set to 0.999 and linearly increases to 0.9999 during the training. We pretrain all the models with the commonly used YFCC15M dataset, which is flited from the YFCC100M [14] dataset by [12].

For the ICinW academic track experiment, we pretrain the model with three datasets: YFCC-15M [14], GCC3M [13]+12M [2] and ImageNet-21K [6] (ImageNet-1K data is excluded). Here we use the UniCL [16] to utilize the ImageNet-22k dataset in the pretraining with a unified format. We train the model for 32 epochs and 16384 batch size, the rest settings are the same as the YFCC15M setting.

Zero-shot ImageNet-1K classification. For zero-shot on ImageNet-1K, we follow the prompt setting in [11] to convert the labels to text features, which contains 7 prompt templates and we use the average feature as the final label feature. We calculate the similarity between image feature and all the label features to get its zero-shot classification result.

Linear-probing ImageNet-1K classification. For linear probing, we fix the backbone and train a new linear classifier for 90 epochs. Following the setting in MAE [7], we add a batch-norm layer without learnable affine parameters before the classifier to avoid adjusting the learning rate for each model. We set the batch size to 16384 and use the LARS [18] optimizer with weight decay 0 and momentum 0.9. The learning rate is set to 6.4 and decays to 0 following the

cosine schedule.

Fine-tuning ImageNet-1K classification. When fine-tuning on the ImageNet-1K dataset, we average pool the output of the last transformer of the encoder and feed it to a softmax-normalized classifier. We fine-tune 100 epochs for all the experiments, the learning rate is warmed up to 0.0006 for 20 epochs and decay to $1e^{-6}$ following the cosine schedule. Similar to recent works, we also apply the layer decayed learning rate used in [1] and we set the decay factor as 0.7. Note that we use the pure ViT architecture, *without* the techniques used in [1], such as layer scale and relative position embedding. The evaluation metric is top-1 validation accuracy of a single 224×224 crop.

Zero-shot Semantic segmentation. Here we follow the setting in DenseCLIP [19] based on the implementation from mmsegmentaion [4]. For ADE20K and MS-COCO, we report the single-scale test result with 512×512 input. For Pascal Context, we use 480×480 input. To avoid the influence of position embedding caused by changing input size, we use sliding inference with 224×224 input and stride 112. To convert the labels to text embedding, we use 85 prompt templates and use the average feature as the final label feature.

ADE20K Semantic segmentation. Here we use: UperNet [15] based on the implementation from mmsegmentaion [4]. For UperNet, we follow the settings in [1] and use AdamW [10] optimizer with initial learning rate $2e^{-4}$, weight decay of 0.05 and batch size of 16 (8 GPUs with 2 images per GPU) for 160K iterations. The learning rate warmups with 1500 iterations at the beginning and decays with a linear decay strategy. We use the layer decay [1] for the backbone and we set it as 0.6. As the ViT architecture outputs features with the same size, here we add four different scale FPNs to scale the feature map into different size. Specifically, we upsample the output feature of the $4th$ block $4\times$, upsample the output feature of the $6th$ block $2\times$, keep the output feature of the $8th$ block unchanged and downsample the output feature of the $12th$ block $2\times$. We use the default augmentation setting in mmsegmentation including random horizontal flipping, random re-scaling (ratio range [0.5, 2.0]) and random photo-metric distortion. All the models are trained with input size 512×512. The stochastic depth is set to 0.1. When it comes to testing, we report single-scale test result.

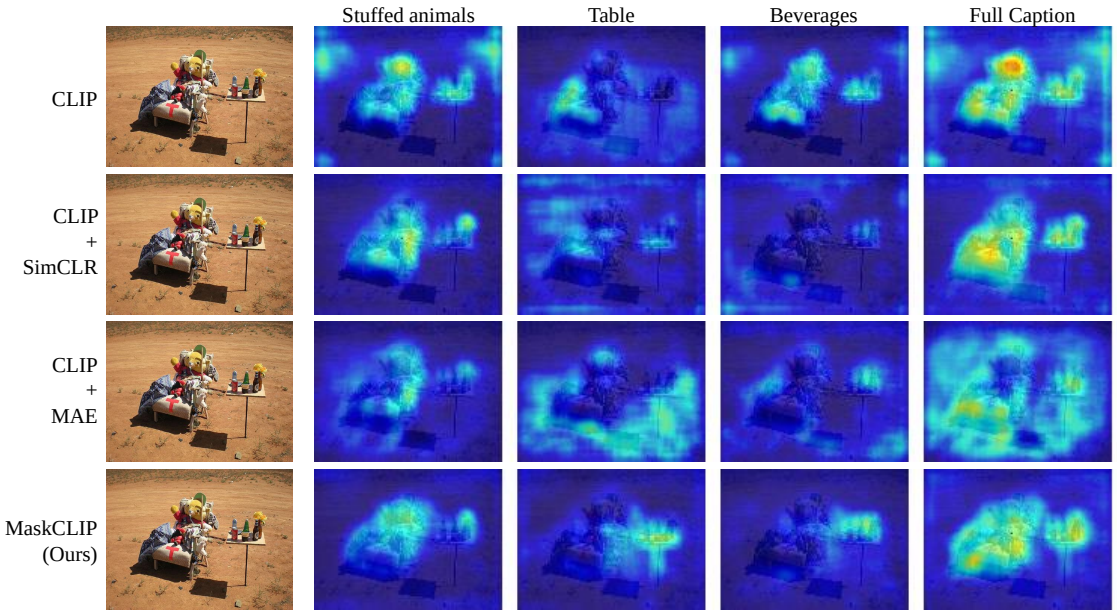
COCO Object Detection and Instance Segmentation. We use the classical object detection framework Mask R-CNN [8] based on the implementation from mmdetection [3]. We train it the $1\times$ schedule with single-scale input (image is resized so that the shorter side is 800 pixels, while the longer side does not exceed 1333 pixels) for 12 epochs. We use AdamW [10] optimizer with a learning rate of $1e^{-4}$, weight decay of 0.05 and batch size of 16. We also use the layer decay [1] for the backbone and we set it as 0.75. The learning rate declines at the $8th$ and $11th$ epoch with decay rate being 0.1. The stochastic depth is set to 0.1. Similar to the implementation of semantic segmentation above, we also use four different scale FPNs to scale the feature map into different size.

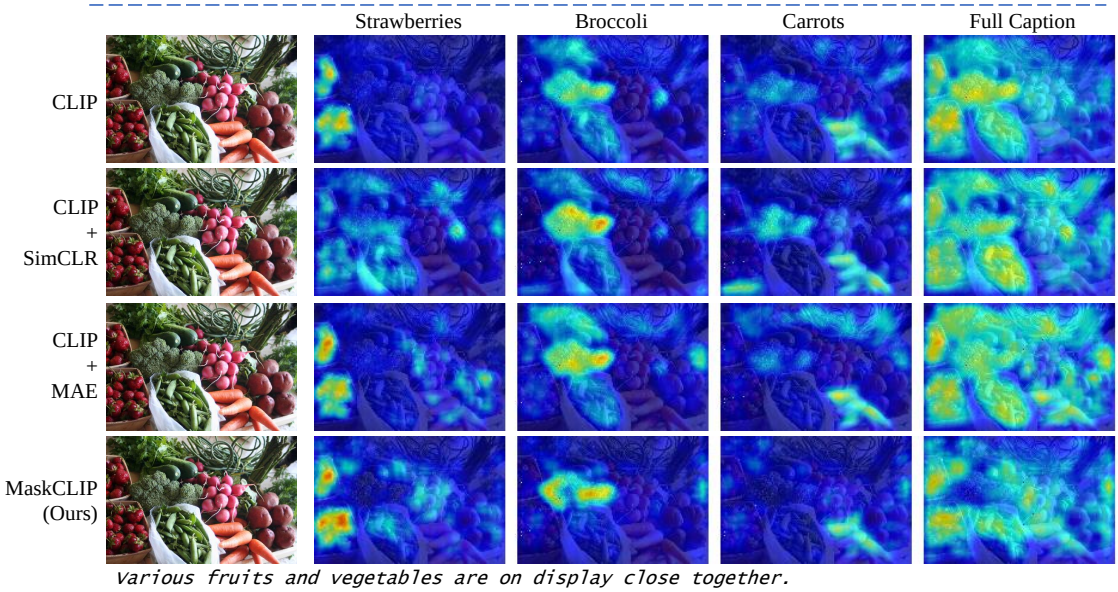
C. More visualization results.

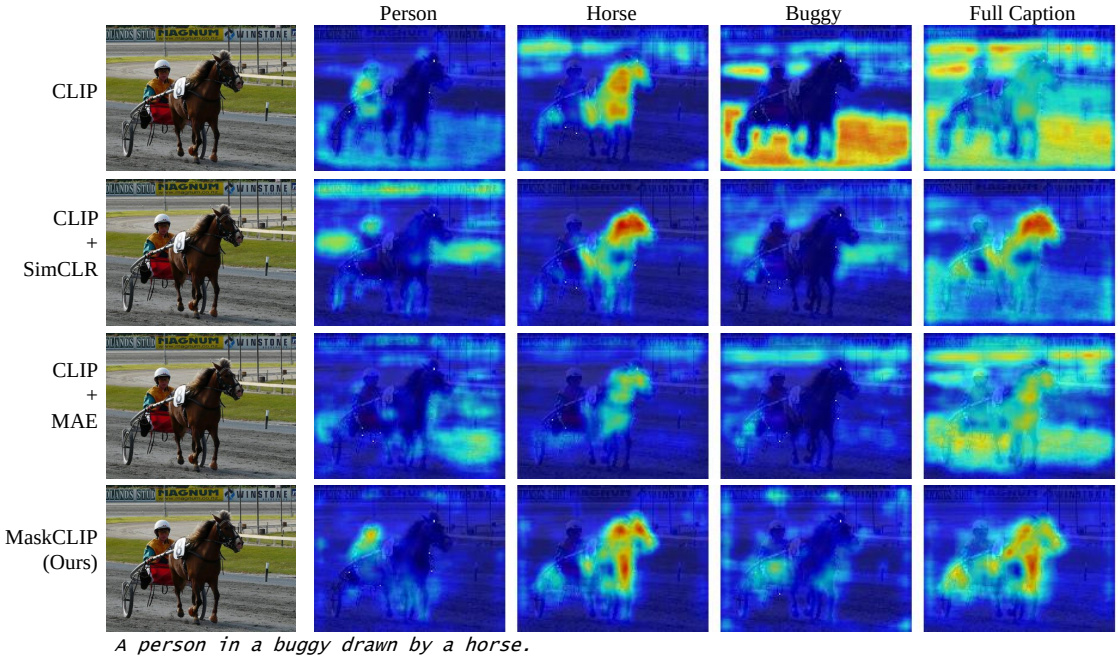
Here we provide more visualization results on the MS-COCO val set. In most cases, our MaskCLIP gets a better feature alignment performance between image and text.

D. Societal impacts

MaskCLIP is an improvement of CLIP, so it has the same societal impacts of CLIP, including some malicious usages and positive applications. Meanwhile, CLIP and MaskCLIP may suffer from some unwanted data bias, as the data used for training are roughly collected from the Internet.


Stuffed animals
Table
Beverages
Full Caption
CLIP
CLIP + SimCLR
CLIP + MAE
MaskCLIP (Ours)
Large stuffed animal posed outdoors as if sitting in a chair with beverages on a table.


Strawberries
Broccoli
Carrots
Full Caption
CLIP
CLIP + SimCLR
CLIP + MAE
MaskCLIP (Ours)
Various fruits and vegetables are on display close together.


Person
Horse
Buggy
Full Caption
CLIP
CLIP + SimCLR
CLIP + MAE
MaskCLIP (Ours)
A person in a buggy drawn by a horse.

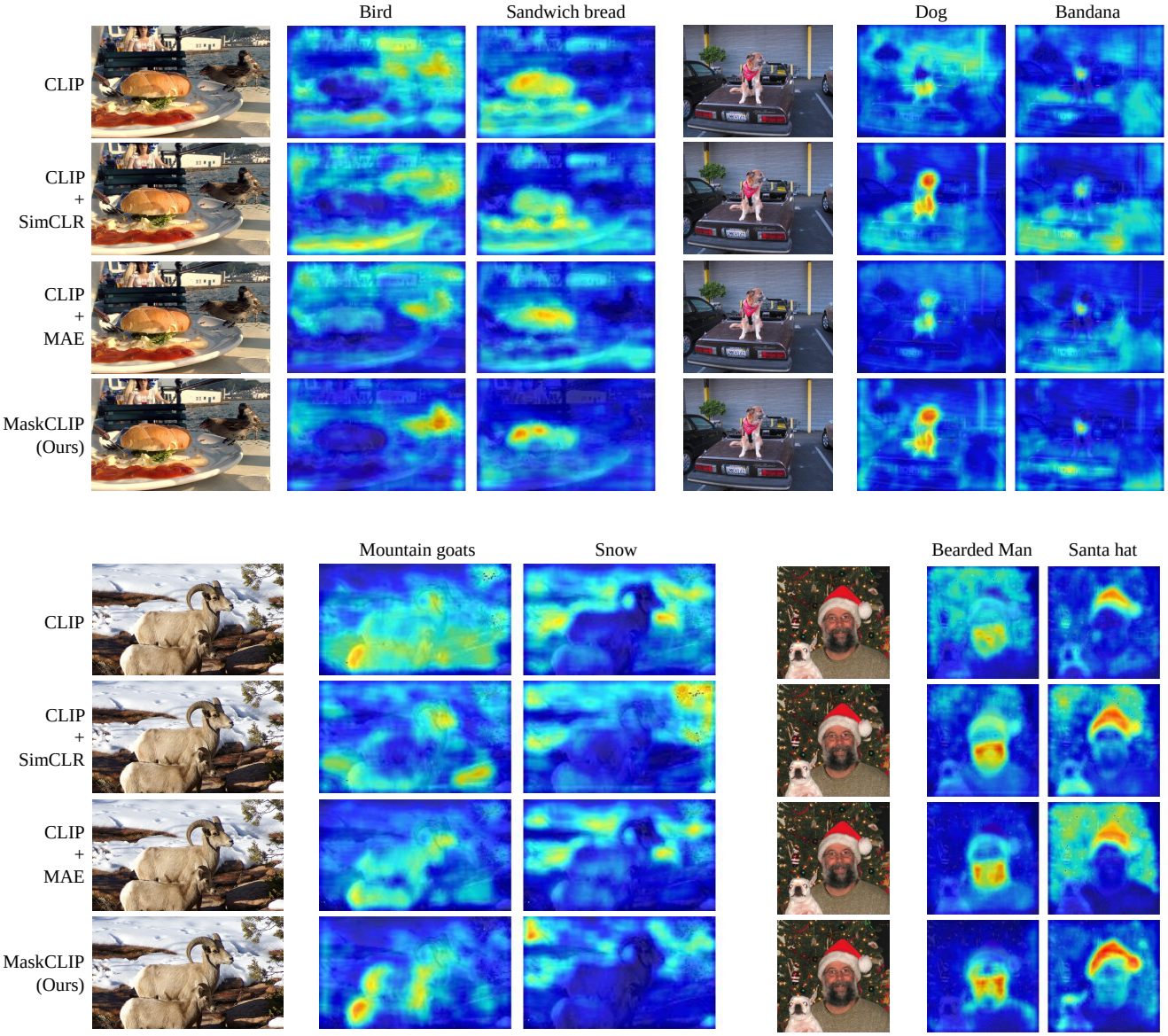


Figure 2. Visualization of the similarity between words and image features. The images and captions are from the MS-COCO val set.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[4] MMSegmentation Contributors. Mmsegmentation, an open source semantic segmentation toolbox. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[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.

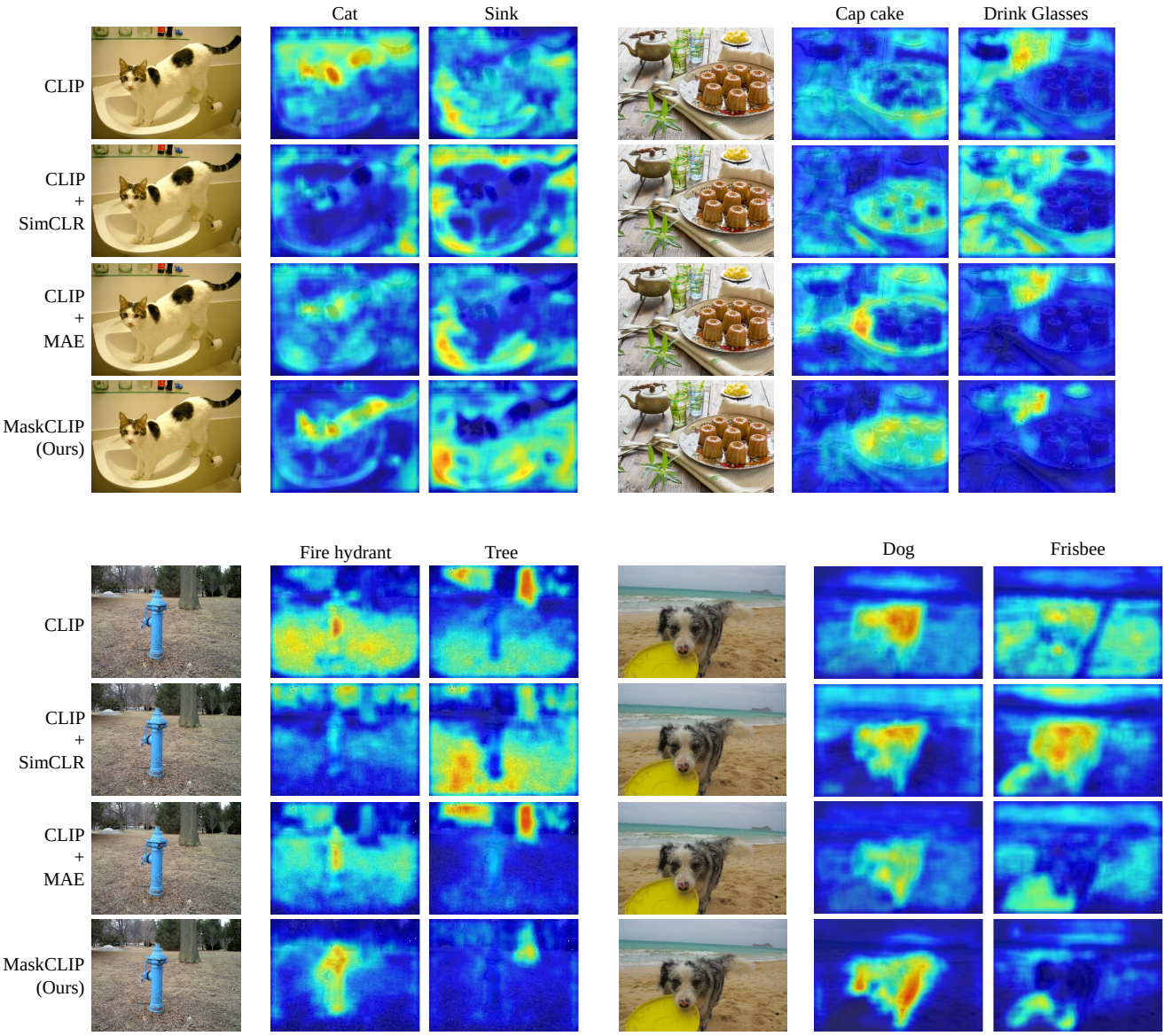


Figure 3. Visualization of the similarity between words and image features. The images and captions are from the MS-COCO val set.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[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 *International Conference on Learning Representations*, 2022.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[11] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[14] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-

Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[15] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[16] 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*, pages 19163–19173, 2022.

[17] 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.

[18] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[19] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021.