Supplementary Materials for
Contrastive Feature Masking Open-Vocabulary Vision Transformer

Dahun Kim Anelia Angelova Weicheng Kuo

Google DeepMind

Appendix

In the supplementary materials, we provide more implementation details and ablations on the hyper-parameters of our CFM-ViT design. We also present more visualizations of the feature reconstruction in our masked image-text pretraining and discuss limitation of CFM-ViT.

A. Implementation Details

Table 1 and 2 summarize the hyperparameters used in our masked image-text pretraining and open-vocabulary detection finetuning, respectively.

B. More Ablations

Our default setting uses the fixed 2D sinusoidal PE in the ViT backbone. In Table 3a, we compare this with the trainable PE and observed no benefits. Note that masked feature reconstruction is *not* added in this experiment.

In Table 3b, we ablate the number of the reconstruction decoder blocks and observe two decoder blocks work the best. Table 3c ablates the loss coefficient between the contrastive and reconstruction losses. We set L_{con} : L_{rec} = 1 : 2 as our default setting.

C. Visualizations: Feature Reconstruction

Fig. 1 provides qualitative examples of the feature reconstruction task in our masked image-text pretraining (Sec. 3.2). We test the pretrained model on the Flickr image-text paired dataset. Our reconstruction branch takes a heavily masked image (middle image per example) and predicts the masked features in the joint image-text embedding space. We visualize the similarity map between the *reconstructed* image features and a query text embedding (right image per example). We observe that the learned reconstructions are semantically plausible with respect to the queried image-text pairs.

configuration	
optimizer	AdamW
momentum	β=0.9
weight decay	1e-2
learning rate	5e-4
warmup steps	1e4
total steps	5e5
batch size	4096 or 16384
image size	224
stochastic depth	0.0
positional embedding (encoder)	fixed 2D sinusoidal
positional embedding (decoder)	fixed 2D sinusoidal
# recon. decoder blocks	2
mask ratio (contr. encoder)	0%
mask ratio (recon. encoder)	75%
loss weights (L_{con} vs L_{rec})	1:2

Table 1: Hyperparameters for CFM-ViT **pretraining**.

configuration	LVIS / COCO
optimizer	SGD
momentum	β=0.9
weight decay	1e-4 / 0.01
learning rate	0.18 / 0.02
backbone lr ratio	0.5×
step decay factor	0.1×
step decay schedule	[0.8, 0.9, 0.95]
warmup steps	1k
total steps	36.8k / 11.3k
batch size	128
image size	1024
stochastic depth	0.0
positional embedding	fixed 2D sinusoidal
α, β in Eq.2	0.65, 0.35

Table 2: Hyperparameters for CFM-ViT **finetuning** on open-vocabulary detection.

D. Limitations

CFM-ViT utilizes the knowledge in pretrained Vision Language Models. The resulting detector model weights will reflect the data biases. In this paper, we mainly demonstrate CFM-ViT's capabilities in comparison with existing open-vocabulary detection works.

	AP$_r$	AP
fixed sinusoidal	27.4	30.4
trainable	27.2	30.3

(a) **Fixed sinusoidal PE vs trainable PE.**

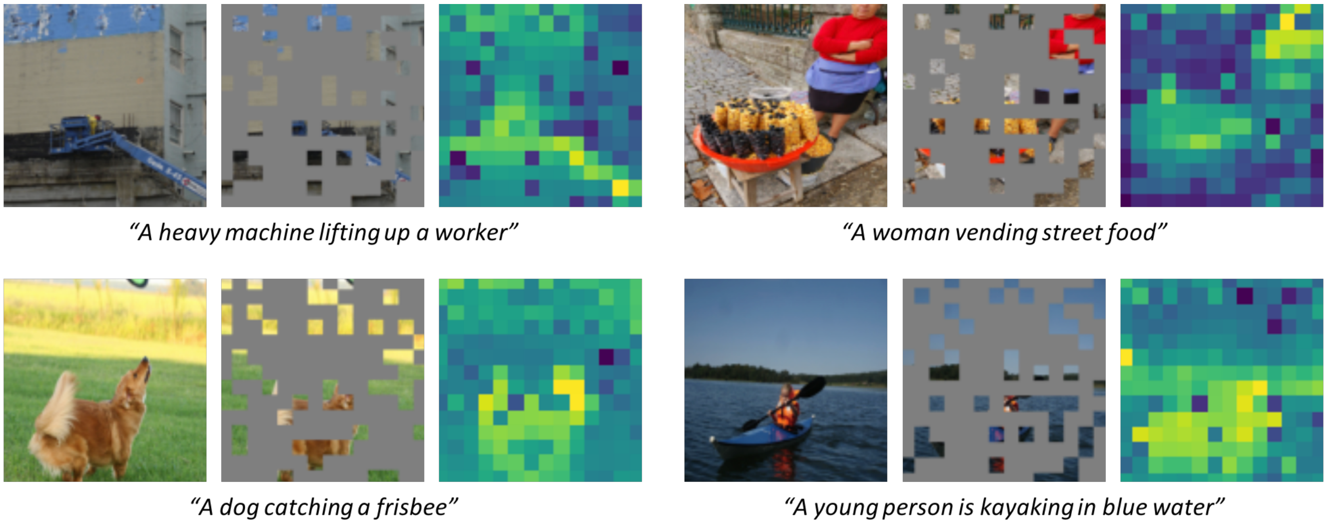
# dec. blocks	AP$_r$	AP
1	30.0	33.7
2	30.7	34.0
4	29.9	34.0

(b) **Number of recon. decoder blocks.**

$L_{con} : L_{rec}$	AP$_r$	AP
1:1	30.2	33.3
1:2	30.7	34.0
1:5	30.3	34.0

(c) **Loss weights between L_{con} and L_{rec}.**

Table 3: **Ablation study** on LVIS open-vocabulary detection benchmark. ViT-L/16 backbone and contrastive batch size 4k are used unless otherwise noted. Note that masked feature reconstruction is *not* used in subtable (a). Our best setting is marked by gray.



"A heavy machine lifting up a worker"

"A woman vending street food"

"A dog catching a frisbee"

"A young person is kayaking in blue water"

Figure 1: **Feature reconstruction visualization.** For each example, we visualize the (left) original image, (middle) masked image, and (right) the similarity map between the *reconstructed* features and the text query embedding (bottom). We observe that our CFM-ViT is able to predict whole-image semantics from the heavily masked image.

E. Dataset License

- COCO [3]: Creative Commons Attribution 4.0 License
- LVIS [2]: CC BY 4.0 + COCO license
- COCO Captions (retrieval) [1]: CC BY
- Flickr30k (retrieval) [4]: Custom (research-only, non-commercial)
- Objects365 [5]: Custom (research-only, non-commercial)

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. In *https://arxiv.org/abs/1504.00325*, 2015. 2

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2

[3] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[4] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 2

[5] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 2