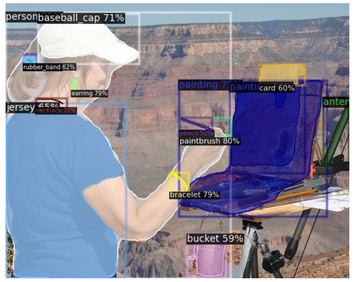
Object Prior Embedded Network for Query-Agnostic Image Retrieval

Yikang Li, Jenhao Hsiao, Chiuman Ho
OPPO US Research Center
{yikang.li1, mark, chiuman}@oppo.co

Motivation

➢ Conventional Object Detection: lack information of the entire image and cannot be applied out-of-the-domain.



"A woman wearing a white hat and a black shirt paints a scene from the Grand Canyon"

➢ "Grand Canyon" will miss based on object detection methods.

➢ "A woman", "white", and "black" will miss based on conventional methos

The Pretrained-VLM will eliminate the first disadvantage, and the open-vocabulary (Detic [1]) will eliminate the second disadvantage.

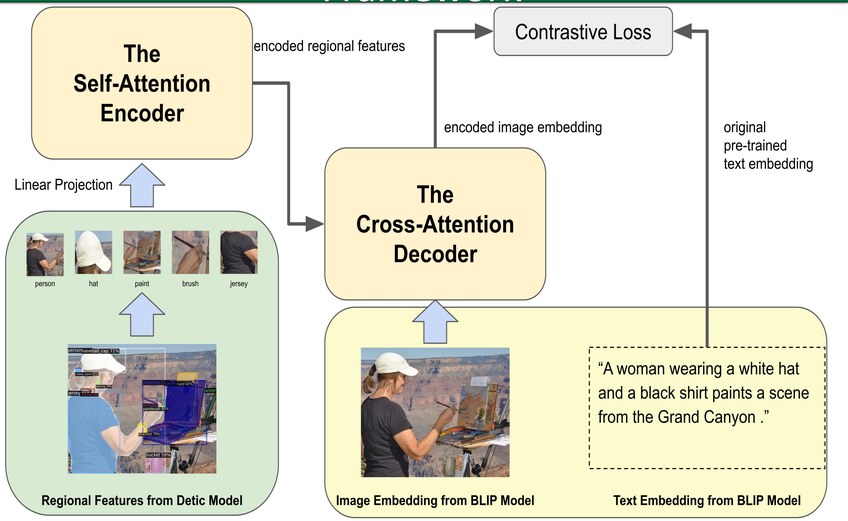
➢ Good VLM is huge and hard to finetune. Transfer learning is a good choice. [2]

1	SEER (RG-10B)	85.8%	10000M
2	V-MoE-H/14 (Every-2)	88.36%	7200M
3	V-MoE-L/16 (Every-2)	87.41%	3400M
4	V-MoE-H/14 (Last-5)	88.23%	2700M
5	CoAtNet-7	90.88%	2440M
6	CoCa (finetuned)	91.00%	2100M

Good VLM require "perfect" hardware and not friendly for finetuning.

✓ **Transfer learning is a good way to use the huge VLM.**

Framework



Method

➢ Region Features Extraction Strategy

$$O_i = \begin{cases} \mathcal{F}_{obj}(image), & \text{if } \frac{Area(obj_i)}{Area(image)} > \alpha \\ \text{skip}, & \text{otherwise} \end{cases}$$

➢ The Cross-Attention Decoder Block

$$q = \mathbf{LN}(I), k = v = \mathbf{LN}(\hat{X})$$

$$Y = I + \mathbf{MHSA}(q, k, v, \mathbf{None})$$

$$\hat{Y} = Y + \mathbf{FFN}(Y)$$

Performance

➢ Performance on Flickr30K

Method name	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1↑	R@5 ↑	R@10↑
SCAN [9]	67.4	90.3	95.8	48.6	77.7	85.2
SGM [20]	71.8	91.7	95.5	53.5	79.6	86.5
CAAN [24]	70.1	91.6	97.2	52.8	79.0	87.9
DPRNN [3]	70.2	91.6	95.8	55.5	81.3	88.2
MMCA [21]	74.2	92.8	96.4	54.8	81.4	87.8
IMRAM [2]	74.1	93.0	96.6	53.9	79.4	87.2
SHAN [8]	74.6	93.5	96.9	55.3	81.3	88.4
SGRAF [5]	77.8	94.1	97.4	58.5	83.0	88.8
MEMBER [12]	77.5	94.7	97.3	59.5	84.8	91.0
DIME [16]	81.0	95.9	98.4	63.6	88.1	93.0
PreTrain-linear-prob	82.3	96.9	98.8	65.1	89.9	93.4
Ours	**88.9**	**98.2**	**99.2**	**73.6**	**92.9**	**96.2**

➢ Inference Speed on MSCOCO 5K

Method	Time (second)
query-dependent	6349.5
query-agnostic	**167.2**

Method name	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1↑	R@5 ↑	R@10↑
SCAN [9]	50.4	82.2	90.0	38.6	69.3	80.4
SGM [20]	50.0	79.3	87.9	35.3	64.9	76.5
CAAN [24]	52.5	83.3	90.9	41.2	70.3	82.9
MMCA [21]	54.0	82.5	90.7	38.7	69.7	80.8
IMRAM [2]	53.7	83.2	91.0	39.7	69.1	79.8
SGRAF [5]	58.8	84.8	92.1	41.6	70.9	81.5
MEMBER [12]	54.5	82.3	90.1	40.9	71.0	81.8
DIME [16]	59.3	85.4	91.9	43.1	73.0	83.1
PreTrain-linear-prob	62.2	86.6	92.7	47.4	75.2	84.4
Ours	**67.8**	**89.0**	**94.2**	**52.4**	**78.5**	**86.7**

➢ Performance on MSCOCO 5K

Contribution

➢ **To our knowledge, we are the first to utilize a cross-attention mechanism to encode the image embeddings as queries and region embeddings as keys and values.**

➢ **The collaboration of the semantic image embeddings and region embeddings (even from different semantic space) can boost the performance of Text-to-Image retrieval task.**

➢ **The experiments also show that a well-defined semantic space is essential for the Text-to-Image retrieval task and the query-agnostic search model is much faster than query-dependent model**

Reference

➢ [1]. Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip KrÅNahenbÅNuhl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision.

➢ [2]. Cited from paper with code website: "https://paperswithcode.com/sota/image-classification-on-imagenet"