ArtVLM: Attribute Recognition Through Vision-Based Prefix Language Modeling

William Yicheng Zhu[1]*, Keren Ye[1]*, Junjie Ke[1], Jiahui Yu[2]†, Leonidas Guibas[1], Peyman Milanfar[1], and Feng Yang[1]

[1] Google Research [2] OpenAI

1 Limitations

One limitation to our proposed method is its increased computational cost. Generative retrieval has n autoregressive text decoding steps, where n is the length of the retrieval template sentence, while contrastive retrieval has one text encoding step. Given the short and fixed-length sentence templates in the attribute learning context, the computational complexity of generative retrieval is $n\times$ constrastive ($n = 2$ to 4). In addition, the text-only attribute embeddings in contrastive retrieval can be precomputed and cached in advance, which would make contrastive retrieval take 0 encoding steps at inference time. This is not possible for generative retrieval, as it is not possible to precompute a part of the likelihood of generating a image-object-attribute triple. Another limitation to the generative retrieval approach is that is is specifically designed for tasks where the assumed lengths of answers or prompts are similar. Since the sum of log probabilities in $L^{(gen)}$ is influenced by the length of the text, the approach is biased towards shorter answers. In the context of attribute prediction tasks, the assumption of similar lengths holds true, allowing us to treat attribute prompt optimization as joint probability optimization in a graph model. This task formulation sets it apart from VQA tasks, which typically involve multiple-choice questions with answers of varying lengths. It is worth noting that this limitation does not undermine our main contribution, which is the development of a novel formulation and framework that connects knowledge from large-scale prefixLM pre-training to the method of generative retrieval for attribute recognition problems.

2 More qualitative examples

We provide more examples to compare our zero-shot retrieval methods, we also include the results from the fully-supervised method SCoNE [14] trained on the VAW dataset. Fig. 1 at the end of the supplementary material shows the results. Some interesting observations can be made. First, VAW is still a closed domain

* Equal contribution.
† Work done at Google.

Table 1: Comparing to the SOTA on the VAW dataset. The top rows show the baseline models; the last three rows shows the results of our method which finetunes the generative prompts. For mA, we report mA@threshold=0.005 as we cross-validated.

Methods	Overall			
	mAP	$mR^{@15}$	mA	$F1^{@15}$
ResNet-Bas.-CE	56.4	55.8	50.3	61.5
LSEP	61.0	50.7	67.1	62.3
PartialBCE+GNN	62.3	52.3	68.9	63.9
ResNet-Bas.	63.0	52.1	68.6	63.9
ML-GCN	63.0	52.8	69.5	64.1
sarafianos2018deep	64.6	51.1	68.3	64.6
SCoNE	68.3	58.3	71.5	70.3
TAP (w/o in-domain PT)	65.4	54.2	67.2	66.4
TAP (in-domain PT)	73.4	63.3	73.5	71.1
Ours*"{A}{O}"*	70.8	61.8	73.7	68.3
Ours*"{O} is {A}"*	72.0	62.1	74.7	68.7
Ours*"{A}{O} is {A}"*	71.9	62.6	74.4	68.7

dataset, lacking in the coverage of long-tailed attributes. In example (2), our generative retrieval predicts "decorative", "antique", and "bamboo", which are visually salient and grammatically correct. However, the ground-truth annotation does not include these two options. Second, compared to others, generative retrieval can surface some of the most significant attributes in the examples. For example, "in the background", "decorative", "worn", or "closed". However, many predictions of the contrastive retrieval method are visually imperceptible or incorrect, such as arch-shaped, standing, partially-eaten, water.

3 Additional Evaluation Results

We include additional results on the VAW experiments in Tab. 1, including the less comparable metrics of $mR^{@15}$ and $F1^{@15}$, which were omitted in the main text due to space constraints. Our method achieves the second place only slightly behind TAP, despite focusing more on cross-domain knowledge extraction and not on constructing task-specific models, which may involve fitting to the evaluation dataset at hand using specialized modules, training procedures, or special training data like segmentation masks that are expensive or impossible to scale.

Furthurmore, to qualitatively demonstrate our model's superior performance on the less frequent categories in the distribution long tail of the Medium (72.0% mAP vs 64.8% mAP) and Tail (60.6% mAP vs 48.0% mAP) attribute classes, we show below Tab. 2 of model performance on the least frequent attributes in VAW:

Table 2: Model performance on the least frequent attributes in VAW

Methods	Model	
	SCoNE mAP	Our mAP
nylon	0.6984	0.5333
bell shaped	0.6955	**0.9167**
braided	0.3893	**0.7046**
styrofoam	0.3591	0.3354
spiral	0.2294	**0.8605**
kissing	0.0409	**0.4085**
wallpapered	0.5293	**0.8956**
smoking	0.1966	**0.3671**
stucco	0.3774	**0.5914**
cubed	0.1102	**0.4258**
TAIL MEAN	0.4800	**0.5940**

4 Image Attribution

In this paper we display several images from the VAW dataset. The Flickr links and the license information for these images can be found in Tab. 3. We thank the original photographers for sharing their photos.

Table 3: Flickr links and license of the images.

Flickr link	User	License
Paper Fig. 4 (from left to right, top to bottom)		
flickr.com/photos/mount_otz/31929683/	mount_otz	CC BY-NC-SA 2.0
flickr.com/photos/jenny-pics/2381135314/	jenny-pics	CC BY 2.0
flickr.com/photos/worldofjan/2984166899/	worldofjan	CC BY-NC 2.0
flickr.com/photos/23909838@N02/3363471858/	23909838@N02	CC BY-SA 2.0
Supplementary materials Fig. 1 (from top to bottom)		
flickr.com/photos/felipelopez/2660779383/	felipelopez	CC BY-NC 2.0
flickr.com/photos/afagen/2269170288/	afagen	CC BY-NC-SA 2.0
flickr.com/photos/nbarcet/2172355975/	nbarcet	CC BY 2.0
flickr.com/photos/dammit_jack/1523816737/	dammit_jack	CC BY-NC 2.0
flickr.com/photos/mjhagen/4347200481/	mjhagen	CC BY 2.0