

Progressive Attention Networks for Visual Attribute Prediction

Supplementary Document

Paul Hongsuck Seo¹
hsseo@postech.ac.kr

Zhe Lin²
zlin@adobe.com

Scott Cohen²
scohen@adobe.com

Xiaohui Shen²
xshen@adobe.com

Bohyung Han³
bhhan@snu.ac.kr

¹ POSTECH
South Korea

² Adobe Research
USA

³ Seoul National University
South Korea

1 Network Architectures on Visual Genome

In PAN, the convolution and pooling layers of VGG-16 network [1], pretrained on ImageNet [2], are used, and three additional attention layers att1, att2 and att3 are stacked on top of the last three pooling layers pool3, pool4 and pool5 respectively as illustrated in Figure 1. The attention functions of att1 and att2 take the local contexts $\mathcal{F}_{i,j}^l$ in addition to the query q and the target feature $f_{i,j}^l$ to obtain the attention score $s_{i,j}^l$. The size of the local contexts is squared with that of the receptive fields of the next two convolution layers before the next attention by setting $\delta = 2$. Two convolutions same as the next two convolution layers in CNN firstly encode the target feature and the local context, and are initialized with the same weights as in CNN (Figure 2a). This embedding is then concatenated with the one-hot query vector and fed to two fully connected layers, one fusing two modalities and the other estimating the attention score. In att3, the attention function takes the concatenation of the query and the target feature and feed it to two fully connected layers (Figure 2b). The attended feature f^{att} obtained from the last attention layer att3 is finally fed to a classification layer to predict the attributes.

The baseline networks also share the same architecture of CNN of VGG-16 network as in PAN (Figure 1). The soft attention and the hard attention is attached to the top of CNN instead in SAN and HAN, respectively. The attention functions in the baselines consist of two fully connected layers taking the concatenation of the query and the target feature as in the attention function of att3 in PAN.

The proposed network and the baselines described above use the query for obtaining the attention probabilities and give us the pure strength of the attention models. However, the target object class, represented by the query, gives much more information than just attention. It

STN-S	STN-M	SAN	HAN	PAN[S]	PAN[H]
conv1_1 (3×3@64)					
conv1_2 (3×3@64)					
pool1 (2×2)					
conv2_1 (3×3@128)					
conv2_2 (3×3@128)					
pool2 (2×2)					
conv3_1 (3×3@256)					
conv3_2 (3×3@256)					
conv3_3 (3×3@256)					
pool3 (2×2)					
↓	att (STN)	↓	↓	att1	att1
conv4_1 (3×3@512)					
conv4_2 (3×3@512)					
conv4_3 (3×3@512)					
pool4 (2×2)					
↓	att (STN)	↓	↓	att2	att2
conv5_1 (3×3@512)					
conv5_2 (3×3@512)					
conv5_3 (3×3@512)					
pool5 (2×2)					
att (STN)	att (STN)	att (soft)	att (hard)	att3 (soft)	att3 (hard)
fc (classification layer)					

Figure 1: Network Architectures of Models.

confines possible attributes and filters irrelevant attributes. For these reasons, we additionally experiment on a set of models that incorporate the target object class conditional prior for the attribute prediction. In these models, the query is fused with the attended feature f^{att} by an additional fully connected layer and the fused feature is used as the input of the classification layer.

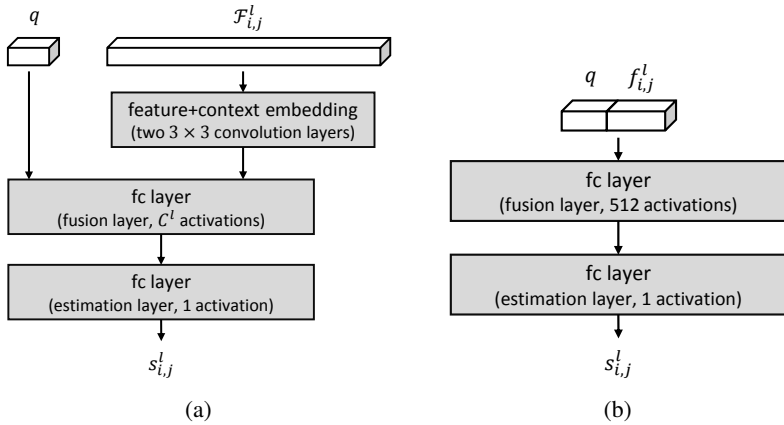


Figure 2: (a) Architecture of the intermediate attention functions $g_{att}^l(\cdot)$ in att1 and att2 of PAN, and (b) architecture of the attention functions of SAN and HAN, and the last attention function of PAN.

2 More Qualitative Results on MBG



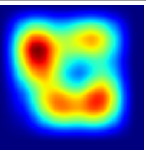
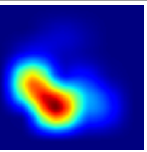
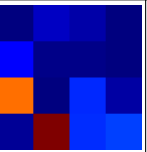


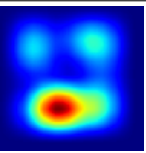
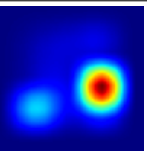
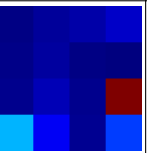


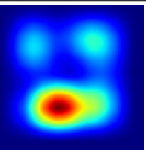
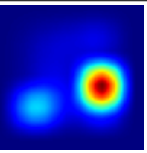
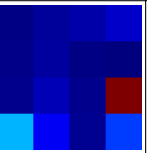
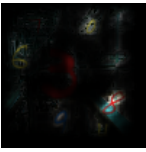
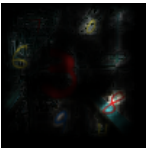
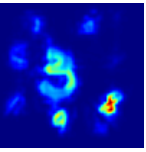
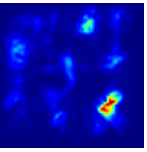
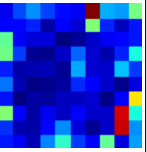
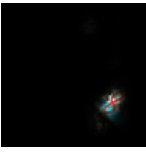
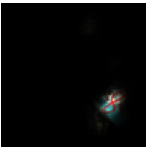
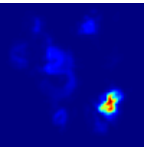
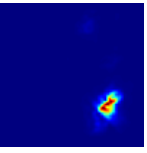
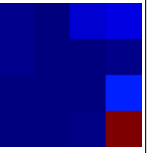
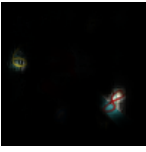
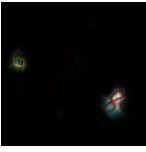
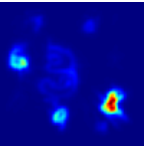
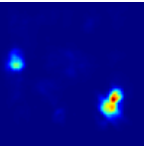
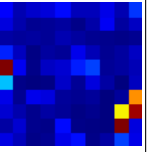
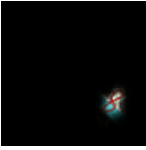
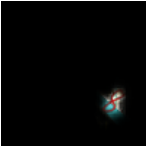
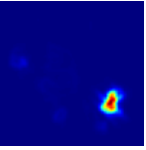
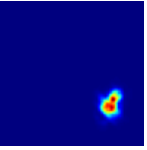
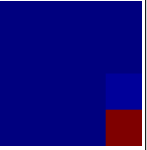
Input image	(a) Masked image	(b) Before attention	(c) After attention	(d) Original resolution	Models
					SAN
					HAN
					HAN
					PAN[S]+CTX (attention 3)
					PAN[S]+CTX (attention 4)
					PAN[H]+CTX (attention 3)
					PAN[H]+CTX (attention 4)

Table 1: Qualitative results of SAN, HAN, PAN[S]+CTX and PAN[H]+CTX with query '8'. (a) Input images faded by attended feature map (c). (b) Magnitude of activations in feature maps $f_{i,j}^l$ before attention; the activations are mapped to original image space by spreading activations to their receptive fields. (c) Magnitude of activations in attended feature maps $\hat{f}_{i,j}^l$ showing the effect of attention in contrast to (b). For PAN[*]+CTX, we only show last two attentions, which accumulate the attentions of earlier layers. Every map is rescaled into $[0, 1]$ by $(x - \min)/(\max - \min)$.


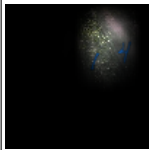
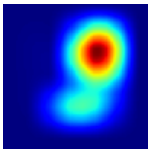
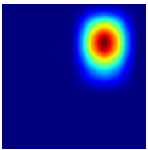

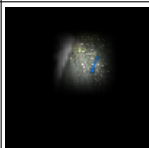
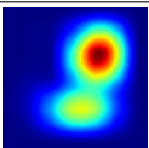
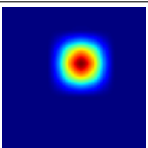
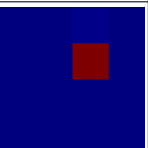
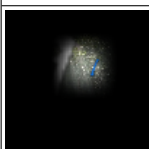
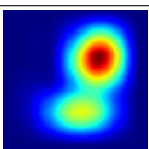
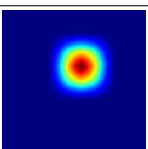
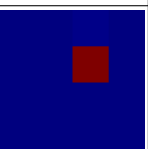
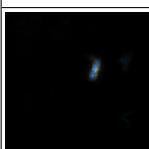
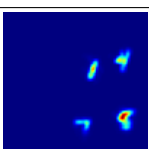
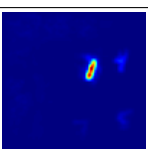
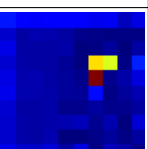
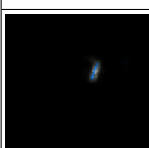
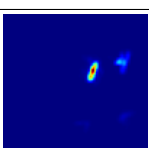
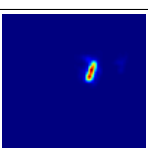

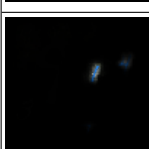
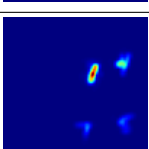
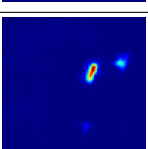
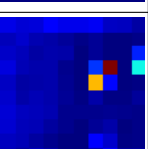

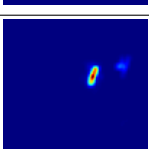
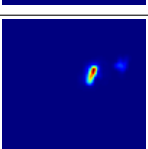
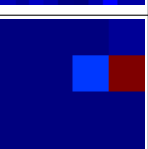
Input image	Masked image	Before att.	After att.	Org. resolution	Models
					SAN
					HAN
					HAN
					PAN[S]+CTX (attention 3)
					PAN[S]+CTX (attention 4)
					PAN[H]+CTX (attention 3)
					PAN[H]+CTX (attention 4)

Table 2: More qualitative results of SAN, HAN, PAN[S]+CTX and PAN[H]+CTX with query '1'.

3 More Qualitative Results on Visual Genome


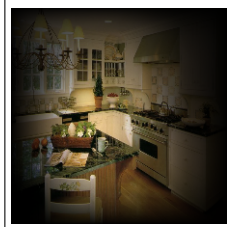
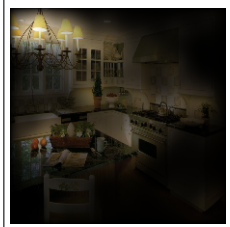
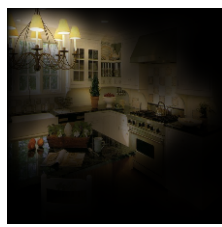
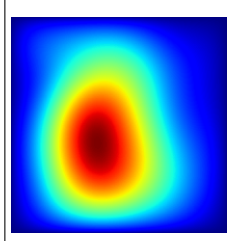
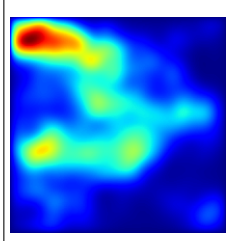
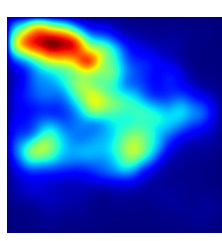

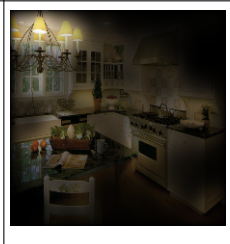
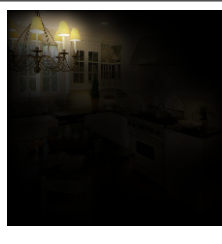
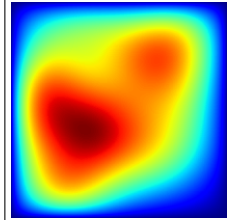
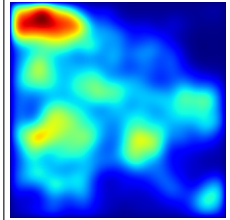
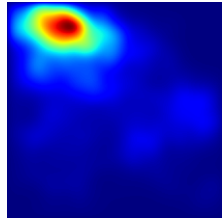
Inputs	SAN	PAN[S]+CTX	
			
Query: light			
	HAN	PAN[H]+CTX	
			
			

Table 3: Attention visualizations of models on VG dataset. Two variants of progressive attention models gradually attend to target objects in fine resolution. For PAN[*]+CTX, we only show last two attentions, which accumulate attentions of earlier layers. More qualitative results are presented in supplementary document.





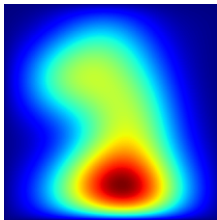
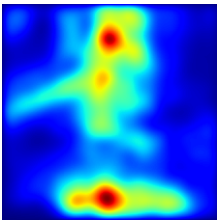
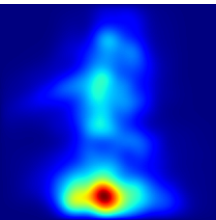



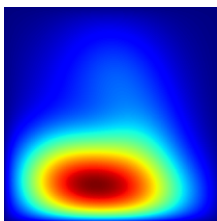
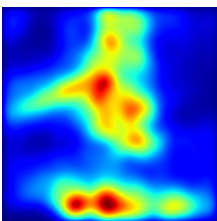
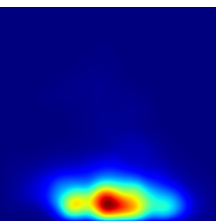
Inputs	SAN	PAN[S]+CTX	
			
<p>Query: cap</p>			
	HAN	PAN[H]+CTX	
			
			

Table 4: More visualizations of attentions.

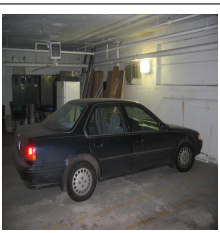
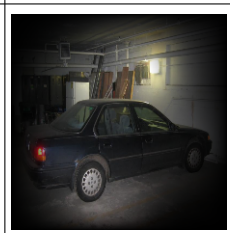
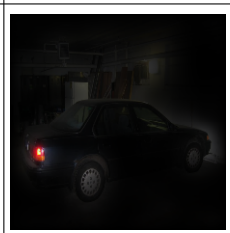
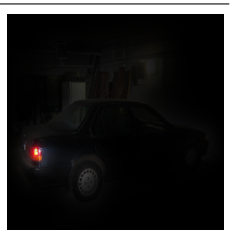
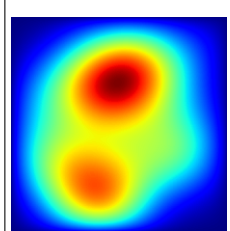
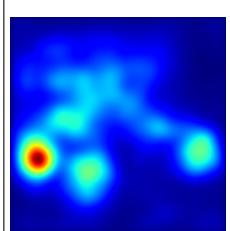
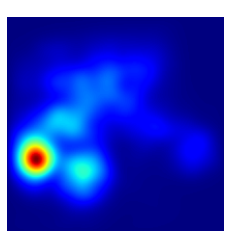
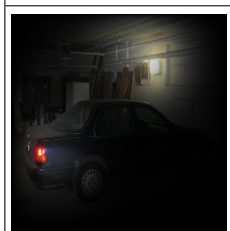
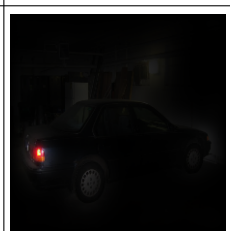
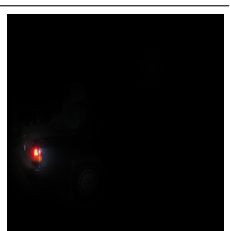
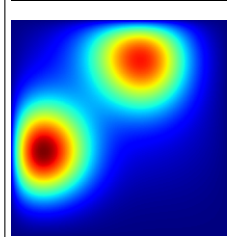
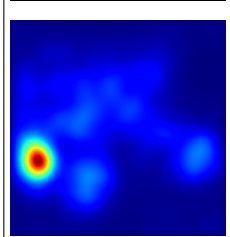
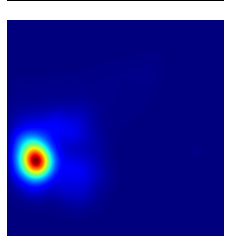
Inputs	SAN	PAN[S]+CTX	
			
Query: light			
	HAN	PAN[H]+CTX	
			
			

Table 5: More visualizations of attentions.

References

- [1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.