PROGRESSIVE ATTENTION NETWORKS FOR VISUAL ATTRIBUTE PREDICTION

Paul Hongsuck Seo[†], **Zhe Lin**[‡], **Scott Cohen**[‡], **Xiaohui Shen**[‡] **& Bohyung Han**[†]
[†]POSTECH, Korea
[‡]Adobe Research
{hsseo, bhhan}@postech.ac.kr
{zlin, scohen, xshen}@adobe.com

ABSTRACT

We propose a novel attention model which can accurately attend to target objects of various scales and shapes in images. The model is trained to gradually suppress irrelevant regions in an input image via a progressive attentive process over multiple layers of a convolutional neural network. The attentive process in each layer determines whether to pass or suppress features at certain spatial locations for use in the next layer. We further employ local contexts to estimate attention probability at each location since it is difficult to infer accurate attention by observing a feature vector from a single location only. The experiments on synthetic and real datasets show that the proposed attention network outperforms traditional attention methods in visual attribute prediction tasks.

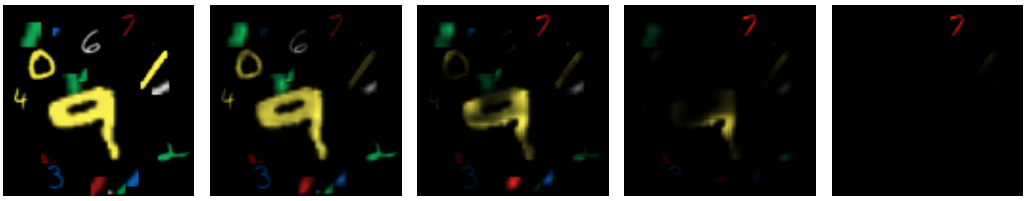
1 INTRODUCTION

Attentive mechanisms often play important roles in modern neural networks (NNs) especially in computer vision tasks. Many visual attention models have been introduced in the previous literature, and they have shown that attaching an attention to NNs can improve the accuracy in various tasks such as image classification (Jaderberg et al., 2015; Ba et al., 2015; Mnih et al., 2014; Larochelle & Hinton, 2010), image generation (Gregor et al., 2015), image caption generation (Xu et al., 2015) and visual question answering (Yang et al., 2015; Andreas et al., 2016; Xu & Saenko, 2015).

There are several motivations for incorporating attentive mechanisms in NNs. One of them is that it is analogous to the perceptual process of human beings. The human visual system concentrates attention to a region of interest instead of processing an entire scene. Likewise, in a neural attention model, we can focus processing only on attended areas of the input image. This benefits us in terms of computational resources; the number of hidden units may be reduced since the hidden activations only need to encode the region with attention (Mnih et al., 2014).

Another important motivation is that some computer vision tasks, *e.g.* visual question answering (VQA), require identifying the object for accurate attribute prediction. For example, when the input image contains multiple objects, the task should focus on the object specified by the question. Figure 1 illustrates an example task to predict the color (answer) of a given input number (query). The query specifies a particular object in the input image (number 7 in this example) for answering its attribute (red). To address this type of tasks, the network architecture should incorporate an attentive mechanism either explicitly or implicitly.

One of the most popular attention mechanisms for NNs is the soft attention method (Xu et al., 2015), which aggregates responses in a feature map weighted by their attention probabilities (see Appendix A for more details). This process results in a single attended feature vector. Since the soft attention method is fully differentiable, the entire network can be trained end-to-end with standard backpropagation. However, it can only model attention to local regions with a certain size depending on the receptive field of the layer chosen for attention. This makes the soft attention method inappropriate for complicated cases, where objects involve significant variations in their scales, and shapes.



(a) input image (b) first attention (c) second attention (d) third attention (e) final attention

Figure 1: An example reference problem (with the query 7 and the answer *red*) and intermediate attention maps using our progressive attention model. It shows that attention is gradually refined through the network layers for resolving the reference problem. Distracting patterns at smaller scales are suppressed at earlier layers while those at larger scales (*e.g.* 9) are suppressed at later layers with larger receptive fields. All attended images are independently rescaled for the visualization.

To overcome this limitation, we propose a novel attention network, referred to as *progressive attention network* (PAN), which enables precise attention over objects of different scales and shapes by attaching attentive mechanisms to multiple layers within a convolutional neural network (CNN). More specifically, the proposed network forces attention prediction in intermediate feature maps by forwarding the attended feature maps in each layer to the subsequent layers in the CNN. Since a feature to be attended in the current feature map is obtained by combining lower-level features with smaller receptive fields, the network can learn to distill the precise spatial support relevant to the target objects as final attention. The contribution of this work is three-fold:

- A novel attention model (progressive attention network) which can be learned to predict attention matching accurate scale and shape of a target object
- Use of local contexts to improve the stability of the progressive attention model
- Achievement of significant performance improvement over traditional soft and hard attention approaches in query-specific visual attribute prediction tasks

The rest of this paper is organized as follows. We first review related work in Section 2. In Section 3, we describe the proposed model with local context information. We then present our experimental results on several datasets in Section 4 and conclude the paper in Section 5.

2 Related Work

Attention on Features The most straightforward attention mechanism is a feature based method, which selects a subset of features by explicitly attaching an attention model to NN architectures. The approaches relying on this attention mechanism have improved performance in many tasks (Xu et al., 2015; Yang et al., 2015; Andreas et al., 2016; Xu & Saenko, 2015; Bahdanau et al., 2015; Luong et al., 2015; Weston et al., 2015; Graves et al., 2014). For example, they have been used to handle sequences of variable lengths in neural machine translation models (Bahdanau et al., 2015; Luong et al., 2015), speech recognition (Chorowski et al., 2014) and handwriting generation (Graves, 2013), and manage memory access mechanisms for memory networks (Weston et al., 2015) and neural turing machines (Graves et al., 2014). When applied to computer vision tasks to resolve reference problems, these models are designed to pay attention to CNN features corresponding to subregions in the input image. Image caption generation and visual question answering are typical examples benefited from this attention mechanism (Xu et al., 2015; Yang et al., 2015; Andreas et al., 2016; Xu & Saenko, 2015).

Attention by Image Transformation Another stream of attention models is based on image transformations. These approaches transform a regular grid and sample from the input image with the transformed grid whose element corresponds to a location in the input image. Ba et al. (2015) and Mnih et al. (2014) transform an input image with predicted translation parameters (t_x and t_y) and a fixed scale factor ($\hat{s} < 1$) for image classification or multiple object recognition. Scale factor is also predicted in (Gregor et al., 2015) for image generation, where the network uses Gaussian filters for sampling. Spatial transformer networks (STNs) predict all six parameters of the affine

transformation matrix, and even extend it to a projective transformation and a 16-point thin plate spline transformation (Jaderberg et al., 2015). Because all these transformations used in (Jaderberg et al., 2015) involve scale factors, STNs are capable of dealing with objects in different sizes. However, STN is limited when there are multiple candidate regions for attention. Our model overcomes this problem by formulating attention as progressive filtering on feature maps instead of assuming objects can be roughly aligned by a single spatial transformation.

Multiple Attention Processes There have been several approaches iteratively performing attentive processes to resolve relations between targets. Yang et al. (2015) iteratively attend to images conditioned on the previous attention states for visual question answering as the objects of interest are often not specified explicitly in questions but implicitly in relational expressions about the target objects. Also, Weston et al. (2015) and Graves et al. (2014) incorporate attention mechanisms to memory cells iteratively to retrieve different values stored in the memory. Our proposed model is similar in spirit of iterative attention but aimed at attending to a single target object via operating on multiple CNN layers progressively, *i.e.*, attention information is predicted progressively from feature maps through multiple layers of CNN to capture the fine shapes of the target object.

In (Jaderberg et al., 2015), the authors also conducted an experiment with a network with multiple transformer layers. However, the attention shapes of STNs are still constrained to the type of transformation regardless of the number of transformers. In contrast, the quality of the attention shapes is improved through progressive attention process in the proposed method. Stollenga et al. (2014) introduced a deep network which manipulates intermediate features of a fixed classifier through channel-wise attention process. Although the channel-wise attention process is used at multiple layers of the network to manipulate the intermediate feature representations, they never explored spatial attention process. More importantly, this method requires to have an accurate pretrained classifier for the target classes prior to learning attention while pretraining a general query-specific attribute classifier is not trivial. It is also notable that both (Jaderberg et al., 2015) and (Stollenga et al., 2014) target simple classification tasks without queries while we aim to tackle the query-specific attribute prediction task where answers from a single input image can be very different depending on the input query.

Training Attention Models The networks with soft attention are fully differentiable and thus trainable end-to-end by backpropagation. Xu et al. (2015) and Zaremba & Sutskever (2015) introduced a stochastic hard attention, where the network explicitly selects a single feature based on the predicted attention probability map. Because the explicit selection (or sampling) procedure is not differentiable, REINFORCE learning rule (Williams, 1992), is used to make networks trainable. Transformation based attention models (Ba et al., 2015; Mnih et al., 2014) are mostly trained by REINFORCE learning rule but STN (Jaderberg et al., 2015) proposed a fully differentiable formulation and made it possible to train end-to-end. Compared to these attention networks, the proposed network is also trainable end-to-end by the standard backpropagation without any extra techniques since every operation within the network is differentiable.

3 Progressive Attention Networks

To overcome the limitation of existing attention models in handling variable object scales and shapes, we propose a progressive attention mechanism. In the proposed model, irrelevant features at different scales are suppressed by attention filtering steps in different CNN layers, and computation is focused on the features corresponding to regions of interest. At each attention layer, the model predicts an attention map given the input query and the current feature map via an attention module, and then the attention maps is multiplied to the feature maps channel-wise to obtain attended feature map. In each layer, each attended feature map is then forwarded to the next layer of the CNN for construction of the following feature map, which is illustrated in Figure 2. This progressive attention process allows us to estimate precise details of attention areas while maintaining deep representations appropriate for high-level inference tasks.

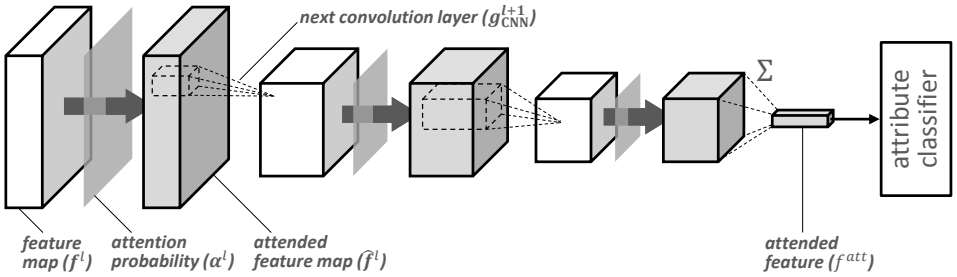


Figure 2: Overall procedure of progressive attention. Attentive processes are repeatedly applied to feature maps at multiple layers and the resulting attended feature maps are used as input feature maps for the next convolution layers in CNN. Attention probabilities α^l are estimated from feature maps and input query. In the last attention layer, the attended feature maps are aggregated to a single feature vector (by sum pooling) and fed to the final attribute classifier.

3.1 Progressive Attentive Process

Let $\boldsymbol{f}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ be an output feature map of a layer $l \in \{0, \dots, L\}$ in CNN with width W_l, height H_l and C_l channels, and $f^l_{i,j} \in \mathbb{R}^{C_l}$ be a feature at (i, j) of the feature map $\boldsymbol{f}^l$. In the proposed PAN, an attentive process is applied to multiple layers of CNN and we obtain the attended feature map $\hat{\boldsymbol{f}}^l = [\hat{f}^l_{i,j}]$, which is given by

$$\hat{f}^l_{i,j} = \alpha^l_{i,j} f^l_{i,j}. \tag{1}$$

Here, the attention probability $\alpha^l_{i,j}$ for a feature $f^l_{i,j}$ is calculated by

$$s^l_{i,j} = g^l_{\text{att}}(f^l_{i,j}, q; \boldsymbol{\theta}^l_{\text{att}}) \quad \text{and} \quad \alpha^l_{i,j} = \begin{cases} \text{softmax}_{i,j}(\boldsymbol{s}^l) & \text{if } l = L \\ \sigma(s^l_{i,j}) & \text{otherwise} \end{cases}, \tag{2}$$

where $g^l_{\text{att}}(\cdot)$ denotes the attention function with a set of parameters $\boldsymbol{\theta}^l_{\text{att}}$ for layer l, $s^l_{i,j}$ is the attention score at (i, j) in layer l, q is the query, and $\sigma(\cdot)$ is a sigmoid function. The attention probability at each location is independent of others in the same feature map, where a sigmoid function is employed to constrain attention probabilities between 0 and 1. For the last layer of attention, we use a softmax function over the entire spatial region for final aggregation of features.

Unlike the soft attention model (see Appendix A), in the intermediate attention layers, the attended feature map $\hat{\boldsymbol{f}}^l$ is not summed up to generate a single vector representation of the attended regions. Instead, the attended feature map is forwarded to the next layer as an input to compute the next feature map, which is given by

$$\boldsymbol{f}^{l+1} = g^{l+1}_{\text{CNN}}(\hat{\boldsymbol{f}}^l; \boldsymbol{\theta}^{l+1}_{\text{CNN}}) \tag{3}$$

where $g^{l+1}_{\text{CNN}}(\cdot)$ is the next CNN operations parameterized by $\boldsymbol{\theta}^l_{\text{CNN}}$.

This feedforward procedure with attentive processes in CNN is repeated from the input of the CNN, *i.e.*, $\boldsymbol{f}^0 = I$, until $\hat{\boldsymbol{f}}^L$ is obtained. Then, the attended feature f^{att} is finally retrieved by summing up all the features in the final attended feature map $\hat{\boldsymbol{f}}^L$ as in soft attention, which is given by

$$f^{\text{att}} = \sum_i^H \sum_j^W \hat{f}^L_{i,j} = \sum_i^H \sum_j^W \alpha^L_{i,j} f^L_{i,j}. \tag{4}$$

The attended feature f^{att} obtained by such process is then used as the input to the visual attribute classifier as illustrated in Figure 2.

In our models, we place the attention layers to the output of max pooling layers instead of every layer in CNN because the reduction of feature resolution within CNN mainly comes from pooling layers. In practice,, we can also skip the first few pooling layers and only attach the attention module to the outputs of last K pooling layers.

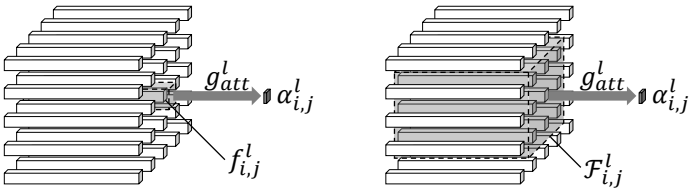


Figure 3: Attention estimation (a) without local context and (b) with local context. In (a), $\alpha^l_{i,j}$ is predicted from $f^l_{i,j}$ only while its spatially adjacent features are also used to estimate $\alpha^l_{i,j}$ in (b).

3.2 Multi-Resolution Attention Estimation

In Eq. (3), the resolution of attention probability map $\boldsymbol{\alpha}^l$ depends on the size of the feature map in the corresponding layer. Due to the nature of a CNN with convolution and pooling layers, the resolution of $\boldsymbol{\alpha}^l$ will decrease with the increasing depth of a layer. Since the attentive processes are performed over multiple layers recursively in our framework, it is possible to attend to the regions of specific sizes and shapes. Note that the proposed network can exploit high-level semantics in deep representations for inference without losing attention resolution.

The progressive attention model is still very effective in predicting fine attention shapes as the attention information is aggregated over multiple layers to suppress irrelevant structures at different granularity. In lower layers, features whose receptive fields contain small distractors are suppressed first. Meanwhile, the features from a part of large distractors remain intact but passed to the next layer delaying its suppression. In higher layers, features of these large distractors would get low attention probability as each feature contains information from larger receptive fields allowing the attention module to distinguish whether the feature is from a distractor or the target object. This phenomenon is well demonstrated in the qualitative results in our experiments (Section 4). An additional benefit of progressive attention is that it is more straightforward during inference since it is a pure feedforward network.

3.3 Local Context

A basic version of PAN discussed so far predicts an attention probability $\alpha^l_{i,j}$ based solely on the feature $f^l_{i,j}$ at a single feature map location. We can improve the quality of attention estimation by allowing the attention layers to observe a local context of the target feature. The local context $\mathcal{F}^l_{i,j}$ of a feature $f^l_{i,j}$ is composed of its spatially adjacent features. For example, the local context can be given by $\mathcal{F}^l_{i,j} = \{f^l_{s,t} | i-\delta \leq s \leq i+\delta, j-\delta \leq t \leq j+\delta\}$ as illustrated in Figure 3. The attention score is now predicted by the attention network with local context as

$$s^l_{i,j} = g^l_{\text{att}}(\mathcal{F}^l_{i,j}, q; \boldsymbol{\theta}^l_{\text{att}}). \tag{5}$$

In this architecture, the area of the local context is given by the filter size corresponding to the composite operation of convolution followed by pooling in the next layer. The local context does not need to be considered in the last layer of attention since its activations are used to compute the final attended feature map. Local context improves attention prediction as it enables the centroid feature to be compared with surrounding features which makes the estimated attention more discriminative.

3.4 Training Progressive Attention Networks

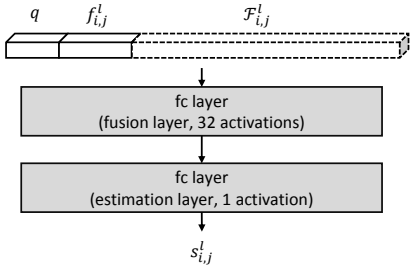
Training a PAN is as simple as training a soft attention network (Xu et al., 2015) because every operation within the network is differentiable. The entire network is trained end-to-end by the standard backpropagation minimizing the binary cross entropies of the object-specific visual attributes. When we train it from a pretrained CNN, the CNN part should always be fine-tuned together since the intermediate attention maps may change the input distributions of their associated layers in CNN.

(a) MREF (b) MDIST (c) MBG

Figure 4: Example of the MREF datasets.

STN	SAN	HAN	PAN
conv1 (3×3@32)			
pool1 (2×2)			
↓	↓	↓	att1
conv2 (3×3@32)			
pool2 (2×2)			
↓	↓	↓	att2
conv3 (3×3@32)			
pool3 (2×2)			
↓	↓	↓	att3
conv4 (3×3@32)			
pool4 (2×2)			
att (STN)	att (soft)	att (hard)	att4
fc (classification layer)			

(a) Network architectures of models on MREF. Arrows represents direct connection to next layer without attention.



(b) Architecture of attention function $g^l_{\text{att}}(\cdot)$. Local contexts $\mathcal{F}^l_{i,j}$ are used only in PAN-CTX.

Figure 5: Detailed illustration of network architectures on MNIST Reference experiments.

4 Experiments

4.1 MNIST Reference

Datasets We conduct experiments on a synthetic dataset created from MNIST (LeCun et al., 1998). The synthetic dataset is referred to as MNIST Reference (MREF; Figure 4a), where each training example is a triple of an image, a query number and its color label. The task on this dataset is to predict the color of the number identified by a query. Five to nine distinct MNIST numbers with different colors in $\{\text{green}, \text{yellow}, \text{white}, \text{red}, \text{blue}\}$ and scales in $[0.5, 3.0]$ are randomly sampled and located in each 100×100 image. When coloring numbers, Gaussian noise is added to the reference color value. To simulate more realistic situations, we made two variants of MREF by chainging backgrounds to either distractors (MDIST; Figure 4b) or natural images (MBG; Figure 4c). Background images in MDIST are constructed with randomly cropped 5×5 patches of MNIST images whereas backgrounds of MBG are filled with natural scene images randomly chosen from the SUN Database (Xiao et al., 2014). The training, validation and test sets contain 30,000, 10,000 and 10,000 images respectively.

Experimental Settings We implement the proposed network with and without the local context observation referred to as PAN-CTX and PAN, respectively. In addition, soft attention network (SAN), hard attention network (HAN) (Xu et al., 2015) and two variants of spatial transformer network (STN-S and STN-M) (Jaderberg et al., 2015), are used as baseline models for comparisons. While STN-S is the model with a single transformer layer, STN-M contains multiple transformer layers in the network. We reimplemented SAN and STNs following the descriptions in (Xu et al., 2015) and (Jaderberg et al., 2015), respectively, and trained HAN by optimizing the marginal log-likelihood loss as it is more accurate and feasible due to small search space in our task. The architecture of image encoding network in SAN and HAN and localization networks in STNs are all identical for fair comparisons. CNN in the proposed network also has the same architecture except for the additional layers for hierarchical attention. The CNN is composed of four stacks of 3×3 convolutions with 32 channels (stride 1) followed by a 2×2 max pooling layer (stride 2) as illustrated in Figure 5a. We used a single fc layer for classification because the task requires simple color prediction. The attention functions $g^l_{\text{att}}(\cdot)$ for all models are formed as multi-layer perceptrons with two layers (Figure 5b).

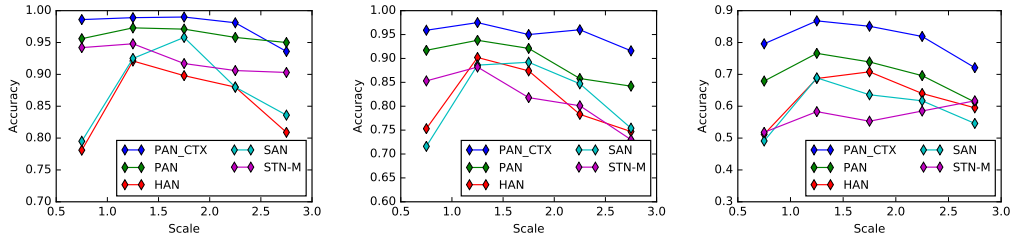
Table 1: Performance of attention models on MREF, MDIST, and MBG datasets.

(a) Color prediction accuracy [%]

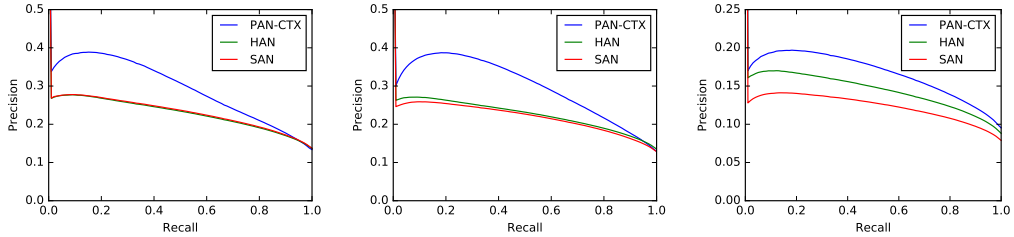
	MREF	MDIST	MBG
STN-S	39.10	38.32	32.27
STN-M	93.89	85.09	52.25
SAN	82.94	75.73	53.77
HAN	81.84	78.49	55.84
PAN	95.92	91.65	69.46
PAN-CTX	**98.51**	**96.02**	**85.55**

(b) True-positive ratio [%]

	MREF	MDIST	MBG
Uniform	2.34	2.35	2.39
SAN	13.61	12.56	6.73
HAN	13.95	13.81	7.64
PAN	17.39	13.10	8.62
PAN-CTX	**22.59**	**22.80**	**11.01**



(a) Attribute prediction accuracies of different models on the test subsets in different scales.



(b) The precision-recall curves of object segmentation with attention probability.

Figure 6: Analysis of algorithms on MREF (left), MDIST (middle), and MBG (right).

The function takes the concatenation of a query q, which is a one-hot vector representing the target object and a feature vector $f^l_{i,j}$, and outputs an attention score $s^l_{i,j}$. In PAN-CTX, the attention functions of $\mathrm{att1}$, $\mathrm{att2}$ and $\mathrm{att3}$ additionally take the local context $\mathcal{F}^l_{i,j}$ containing the adjacent features with $\delta = 2$. Every model is trained from scratch.

Results Table 1a presents color prediction accuracy of all compared algorithms. It is obvious that PAN outperforms all the previous approaches with significant margins and PAN-CTX further improves the performance by exploiting the local contexts for attention estimation. While STN-S often fails to predict the correct answers, STN-M learns to predict the color of the target object through multiple transformations and shows comparable performance to PAN in MREF. However, the performance of STN-M dramatically drops as the dataset becomes more complex and realistic, resulting in even lower performance than SAN and HAN. Also, note that STN-S is capable of attending to any region attended by STN-M since both models predict attention regions by estimating an affine transformation. STN-M achieves the improvement by learning multiple transformers from gradients coming from different levels of features. In contrast to those parametric models, the proposed network can predict attention map with more fine-grained shapes capturing the spatial support of the target object better.

To evaluate the scale sensitivity of each model, we divided the test images into five subsets based on target object scales with uniform interval and computed the accuracies of the models. The results are presented in Figure 6a, where SAN and HAN tend to predict the correct answers only in a scale range between 1.0 and 2.0, while their performance is degraded significantly with wild scale changes. STN-M becomes vulnerable to scale variations in more realistic settings. In contrast, PAN and PAN-CTX are robust to scale variations due to their multi-scale attention machanism especially when the local contexts are incorporated.

Unlike STNs whose attention is constrained to rhombic regions, those models based on feature-wise attention maps can produce attention regions adaptive to the shapes of the target object. We evaluate the attention quality of these models using two complementary criteria: true-positive ratio (TPR)

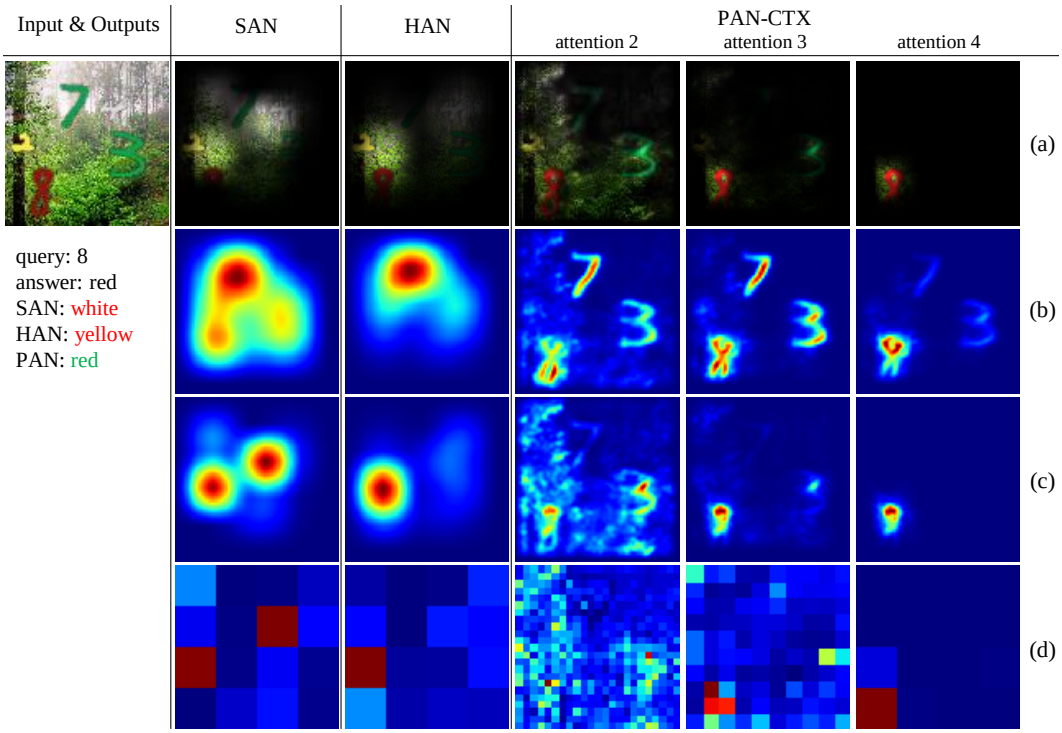


Figure 7: Qualitative results of SAN, HAN and PAN-CTX. (a) Input images faded by attended feature map (c). (b) Magnitude of activations in feature maps $f^l_{i,j}$ before attention: the activations are mapped to original image space by spreading activations to their receptive fields. (c) Magnitude of activations in attended feature maps $\hat{f}^l_{i,j}$ which shows the effect of attention in contrast to (b). (d) Magnitude of activations of the attended feature maps $\hat{f}^l_{i,j}$ in its original resolution of the feature map. For PAN-CTX, only last three attention layers are visualized and attentions of ealier layers are accumulated for visualizing higher attention layers. For HAN, (c) and (d) represent attention probability because attended feature map is not available. Every image except for input image is rescaled into $[0, 1]$ by $(x - min)/(max - min)$.

and precision-recall (PR) curve. TPR measures how strong attention is given to proper location by computing the ratio of the aggregated attention probability within the desired area (a.k.a., ground-truth segmentation) to the attention probability in the whole image (Table 1b). PR measures the overlaps between ground-truth segmentations and binarized segmentation predictions constructed with different thresholds (Figure 6b). Note that the proposed model with the local context observation gives the best results with significant margin compared to all the other methods in terms of both criteria. These results suggest that PAN-CTX constructs more accurate shapes of attended regions than all other attention models.

Figure 7 shows the qualitative results of the proposed method and two baselines on the MBG dataset. The proposed model yields accurate attention regions eventually by gradually augmenting attention and suppressing irrelevant regions in the image. We can observe that the proposed model could maintain the high attention resolution through the progressive attention process. In contrast, the baseline models attend to the target objects only once at the top layer resulting in a coarse attention in size and shape. More qualitative results in these experiments are presented in Appendix C.

4.2 Attribute Prediction on Visual Genome

Dataset Visual Genome (VG) (Krishna et al., 2016) is an image dataset containing several types of annotations: question/answer pairs, image captions, objects, object attributes and object relationship. We formulate the object attribute prediction as a multi-label classification task with reference. Given an input image and a query (*i.e.*, an object category), we predict the binary attributes of individual objects specified by a query. We used 827 object classes and 749 attribute classes that appear more

Table 2: Weighted mAP of the attribute prediction and TPR of attentions measured with ground-truth bounding boxes on VG dataset.

	attention only		w/ prior	
	mAP	TPR	mAP	TPR
SAN	27.62	15.01	31.84	17.65
HAN	27.72	17.24	31.93	19.70
PAN-CTX	**29.38**	**18.01**	**32.50**	**20.17**

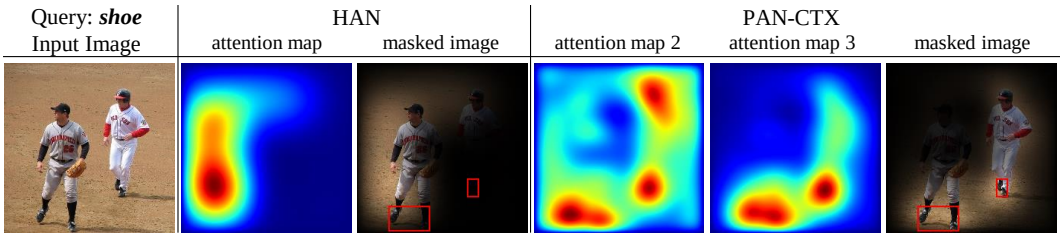


Figure 8: Visualization of example attentions of HAN and PAN-CTX on VG dataset. Attention maps present magnitude of attended features and red boxes show ground truth bounding boxes of query.

than 100 times. A total of 86,674 images with 667,882 object attribute labels are used for our experiment, and they are split into training, validation and test sets each containing 43,337, 8,667 and 34,670 images. The task is challenging because scales of objects largely vary and the attributes may be associated with very small objects.

Experimental Settings and Results We mainly compare our algorithm with SAN and HAN since STNs could not learn a proper attention process on VG. The transformer layers of STNs generated padded images of different sizes and rotations to encode the query vector to fit the query-specific biases. All the networks share the same CNN architecture of VGG-16 network (Simonyan & Zisserman, 2015), which is pretrained on ImageNet (Deng et al., 2009) and is further fine-tuned on the VG dataset for the attribute prediction. For SAN and HAN, an attention layer is attached to the last pooling layer in VGG-16 while PAN stacks an additional attention layer with the local contexts $\mathcal{F}^l_{i,j}$ with $\delta = 2$ on top of each of the last three pooling layers in VGG-16. We skip to place attention layers at the first two pooling layers ($\mathrm{pool1}$ and $\mathrm{pool2}$) because the features in those layers are not discriminative enough to filter out.We also test models with object class conditional prior. In these models, the final attended feature is fused with the query once more by a fully connected layer allowing the network to reflect the conditional distribution of the attributes given the query. Refer to Appendix B for more detailed descriptions on the network architectures.

All three models are evaluated in terms of mean average precision (mAP) weighted by the frequencies of the attribute labels in the test set, where the computation of mAP follows PASCAL VOC protocol (Everingham et al., 2010). The proposed method consistently achieves the best weighted mAP scores in both experimental settings as shown in Table 2 but the gain reduces with object class conditional prior. Table 2 also shows TPR of each model measured with the ground-truth bounding box for evaluating the attention qualities, and the proposed method shows the best TPR. Figure 8 presents the qualitative results of the proposed network and HAN on VG dataset.

5 Conclusion

We proposed a novel hierarchical attention network, which progressively attends to regions of interest through multiple layers of a CNN. As the model is recursively applied to multiple layers of CNN with an inherent feature hierarchy, it accurately predicts regions of interest with variable sizes and shapes. We also incorporate local contexts into our attention network for more robust estimation. The proposed network can be trained end-to-end with standard error backpropagation. We tested the model on both synthetic and real datasets, and demonstrated significant performance improvement over existing attention methods.

REFERENCES

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. In *CVPR*, 2016.

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: first results. *arXiv preprint arXiv:1412.1602*, 2014.

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.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, pp. 1462–1471, 2015.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pp. 2008–2016, 2015.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *NIPS*, pp. 1243–1251, 2010.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, pp. 2204–2212, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Marijn F Stollenga, Jonathan Masci, Faustino Gomez, and Jürgen Schmidhuber. Deep networks with internal selective attention through feedback connections. In *Advances in Neural Information Processing Systems*, pp. 3545–3553, 2014.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, pp. 1–20, 2014.

Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv preprint arXiv:1511.05234*, 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015.

Wojciech Zaremba and Ilya Sutskever. Reinforcement learning neural turing machines. *arXiv preprint arXiv:1505.00521*, 2015.

Appendices

A Soft Attention Model

In this appendix section, we explain the soft attention network which is introduced in (Xu et al., 2015) and used as one of the baseline models in the experiments. Given a feature map, the soft attention network calculates an attention probability map and uses it to compute the attended feature for classification or other tasks. Given a feature map $\boldsymbol{f} \in \mathbb{R}^{H \times W \times C}$ and a query q containing information of where to attend, a soft attention model first obtains an attended feature map $\hat{\boldsymbol{f}} \in \mathbb{R}^{H \times W \times C}$, where W is width, H is height, and C is the number of channels. The input feature map $\boldsymbol{f}$ is generally a CNN output of an input image I, which is given by

$$\boldsymbol{f} = \mathrm{CNN}(I). \tag{6}$$

For each feature $f_{i,j} \in \mathbb{R}^C$ at (i, j) of the feature map $\boldsymbol{f}$ and the query q, the attention probability map denoted by $\boldsymbol{\alpha} = [\alpha_{i,j}]$ is given by

$$s_{i,j} = g_{\mathrm{att}}(f_{i,j}, q; \boldsymbol{\theta}_{\mathrm{att}}) \tag{7}$$

$$\alpha_{i,j} = \mathrm{softmax}_{i,j}(\boldsymbol{s}), \quad 0 \leq \alpha_{i,j} \leq 1 \tag{8}$$

where $g_{\mathrm{att}}(\cdot)$ is the attention network parameterized by $\boldsymbol{\theta}_{\mathrm{att}}$ and $\boldsymbol{s} = [s_{i,j}]$ is an attention score map. The attention score map is normalized with softmax to produce attention probabilities $\alpha_{i,j}$. Note that $g_{\mathrm{att}}(\cdot)$ can be any kind of network such as a multilayer perceptron.

Let $\hat{f}_{i,j} \in \mathbb{R}^C$ be a vector of the attended feature map $\hat{\boldsymbol{f}}$ at (i, j). Then, the attended feature denoted by $f^{\mathrm{att}} \in \mathbb{R}^C$ is computed by a weighted sum of features as

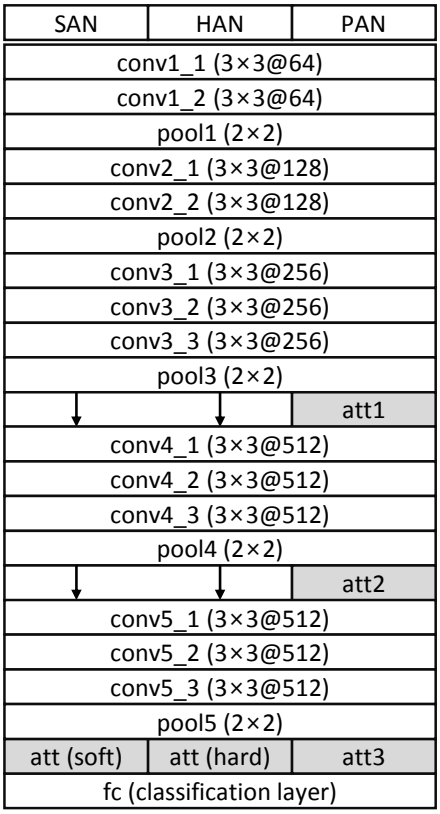
$$f^{\mathrm{att}} = \sum_{i}^{H} \sum_{j}^{W} \hat{f}_{i,j} = \sum_{i}^{H} \sum_{j}^{W} \alpha_{i,j} f_{i,j}. \tag{9}$$

Ideally, the locations in the feature map corresponding to the receptive fields containing an object of interest should have the maximum attention probability while the others have zero probabilities similarly to the hard attention. This statement stands true only if the target object is perfectly aligned with the receptive fields in terms of position and scale. In practice, however, object location and size vary whereas the structure of receptive fields is fixed. Note that there exists the trade-off between the attention resolution and the representation power. If we choose to extract deep and high-level features, we give up high resolution in attention. On the other hand, we need to rely on shallow representations to increase attention resolution. This trade-off limits the performance of existing attention models.

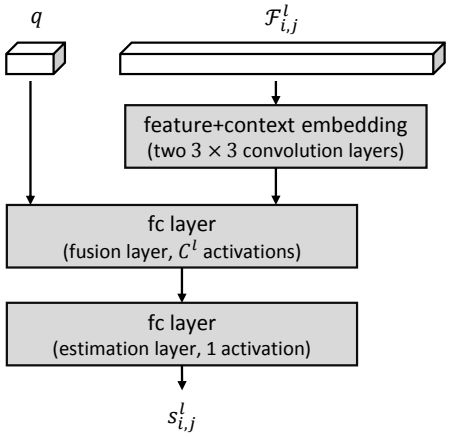
B Network Architectures on Visual Genome

In PAN, the convolution and pooling layers of VGG-16 network (Simonyan & Zisserman, 2015), pretrained on ImageNet (Deng et al., 2009), are used, and three additional attention layers $\mathrm{att1}$, $\mathrm{att2}$ and $\mathrm{att3}$ are stacked on top of the last three pooling layers $\mathrm{pool3}$, $\mathrm{pool4}$ and $\mathrm{pool5}$ respectively as illustrated in Figure 9a. The attention functions of $\mathrm{att1}$ and $\mathrm{att2}$ take the local contexts $\mathcal{F}^l_{i,j}$ in addition to the query q and the target feature $f^l_{i,j}$ to obtain the attention score $s^l_{i,j}$. The size of the local contexts is squared with that of the receptive fields of the next three convolution layers before the next attention by setting $\delta = 3$. Three convolutions same as the next three convolution layers in CNN firstly encode the target feature and the local context, and are initiallized with the same weights as in CNN (Figure 9b). 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 $\mathrm{att3}$, the attention function takes the concatenation of the query and the target feature and feed it to two fully connected layers (Figure 9c). The attended feature f^{att} obtained from the last attention layer $\mathrm{att3}$ is finally fed to a classification layer to predict the attributes.

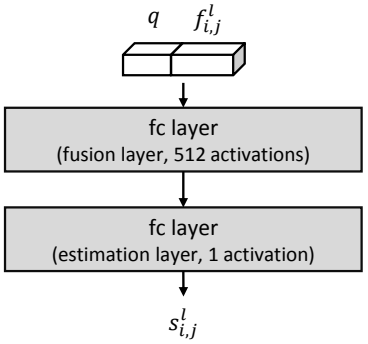
The baseline networks also share the same architecture of CNN of VGG-16 network as in PAN (Figure 9a). In SAN, the soft attention described in Appendix A is attached to the top of CNN. In HAN, the hard attention (Xu et al., 2015) is attached to the top of CNN instead. The hard attention is



(a) Network Architectures of Models.



(b) Architecture of the intermediate attention functions $g^l_{\mathrm{att}}(\cdot)$ in $\mathrm{att1}$ and $\mathrm{att2}$ of PAN.



(c) Architecture of the attention functions of SAN and HAN, and the last attention function of PAN.

Figure 9: Detailed illustration of network architectures on Visual Genome experiments.

implemented to maximize the marginal likelihood directly during training while the original paper maximized the variational lower bound of the marginal likelihood because of the large attention search space. For testing, we also directly calculated the marginal likelihood instead of picking a single prediction with the highest attention probability. This is possible because of relatively small search space of attention in our problem compared to the image captioning where the search space of attention increases exponentially depending on the lengths of sequences. 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 $\mathrm{att3}$ 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 attetion. It 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.

C MORE QUALITATIVE RESULTS ON MNIST REFERENCE

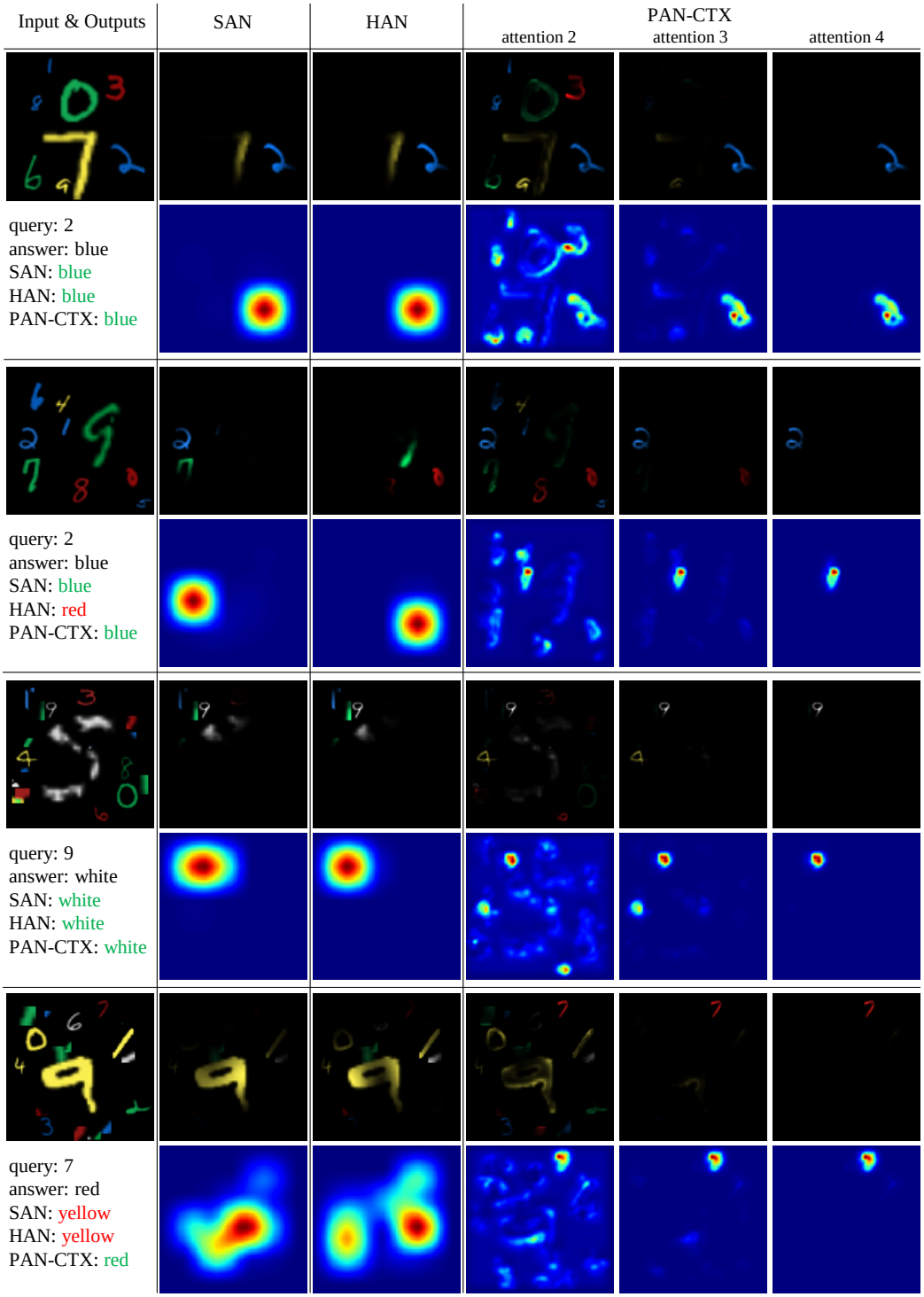


Figure 10: The qualitative results of SAN, HAN and PAN-CTX on the MREF and MDIST datasets. For each example, attended images are shown in the first row and the corresponding attention maps are shown on the second row. In case of the progressive attention network, the last three attention maps (attention 2, 3 and 4) are visualized. As can be seen, attention map at deeper layers reveal the evidence of aggregation over earlier attention maps.

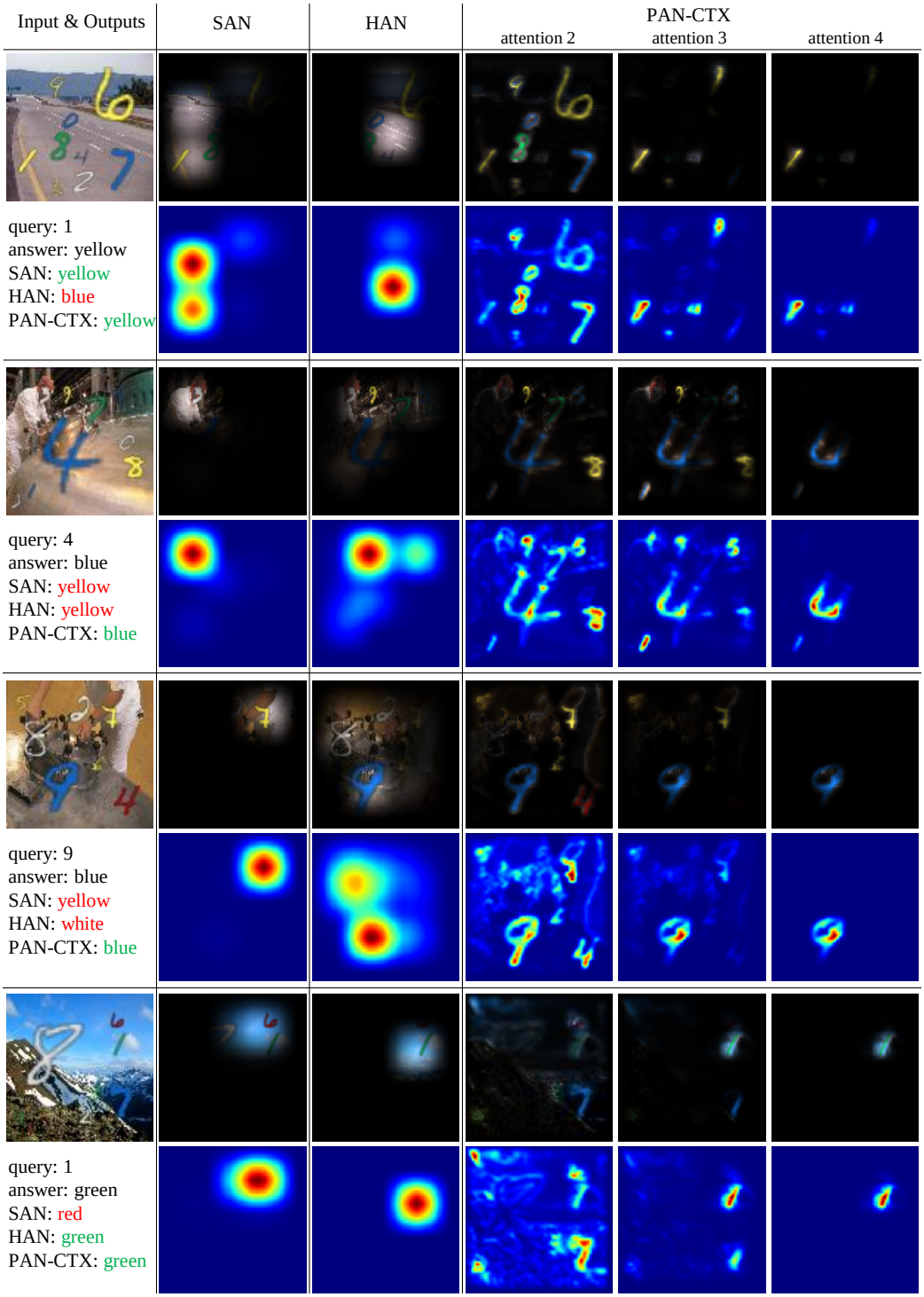
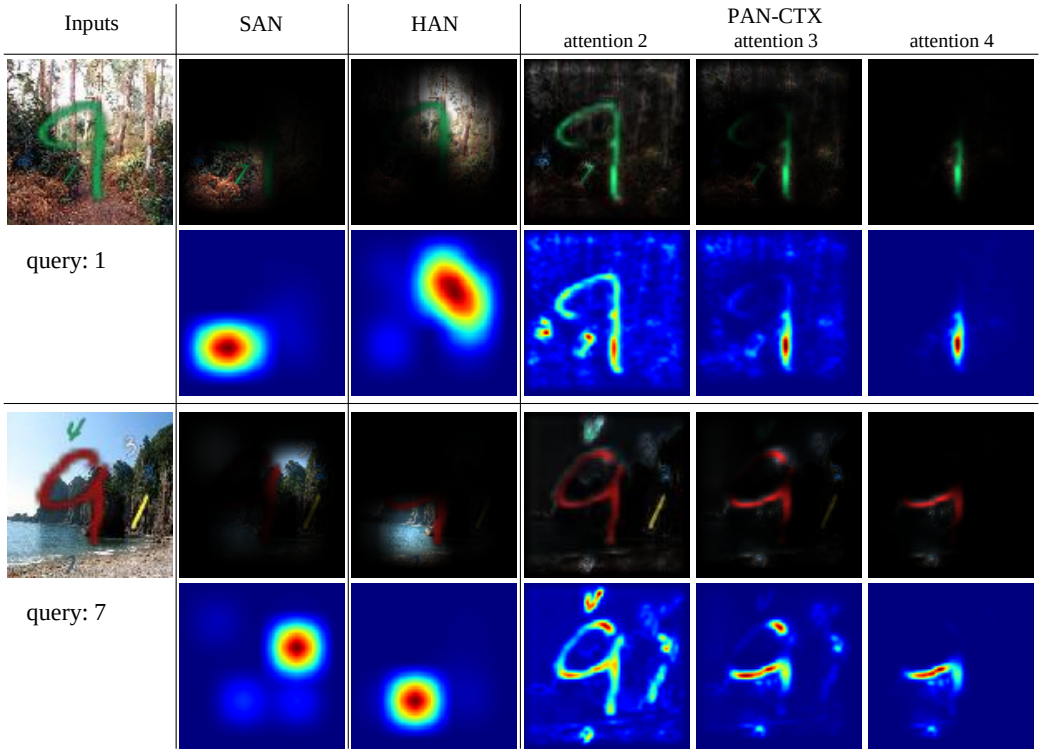
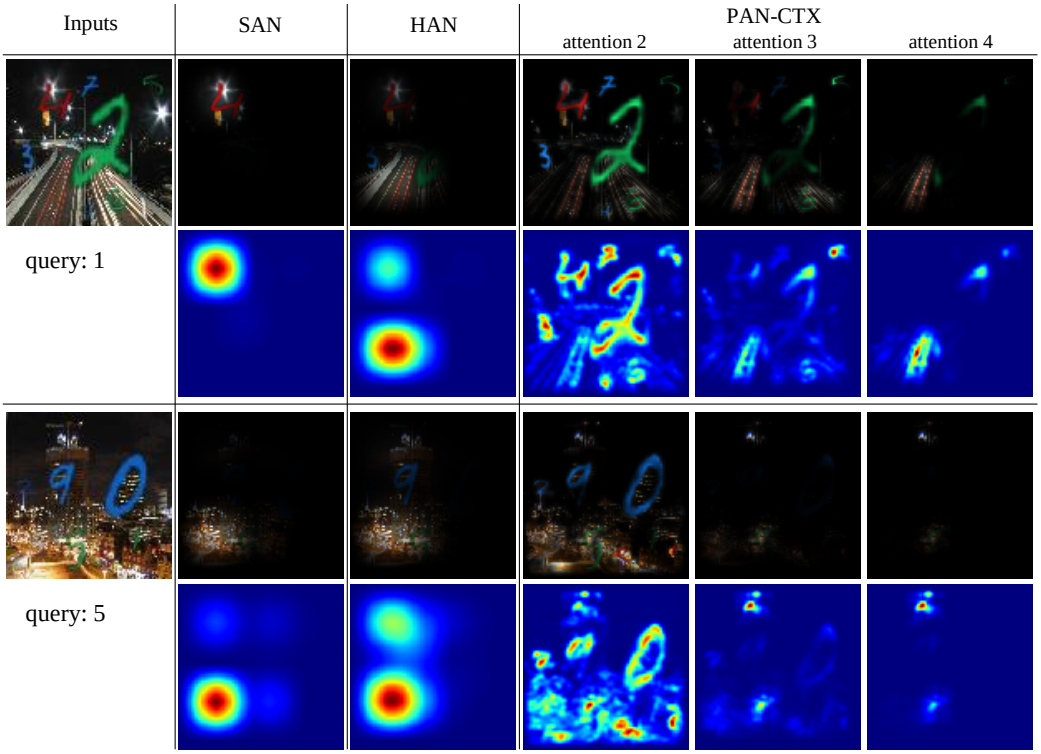


Figure 11: More qualitative results of SAN, HAN and PAN-CTX on the MBG dataset.



(a)



(b)

Figure 12: Two common failure cases of attention models on the MBG dataset. (a) The models attend to a part of a larger structure which resembles the target object. (b) The models are confused by background distractors that are similar to the target object. Although failed, the examples show that the results of PAN-CTX are more visually interpretable (attended to query-like structures).

D MORE QUALITATIVE RESULTS ON VISUAL GENOME

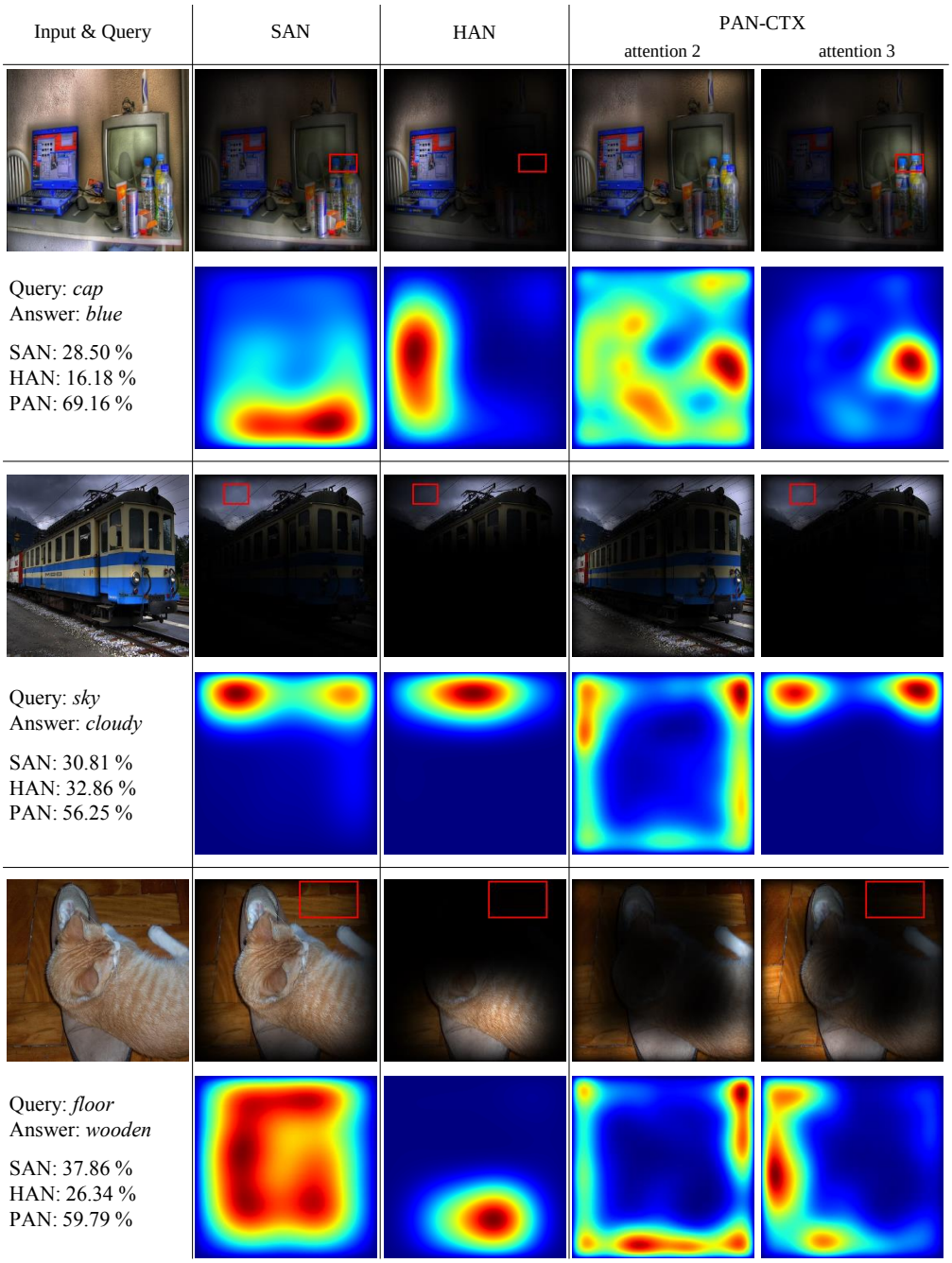


Figure 13: The qualitative results of SAN, HAN and PAN-CTX on the VG dataset. For each example, the attended images are presented in the first row while their attended feature maps are shown in the second row. In the case of the PAN, last two attention maps are visualized where the attention maps at deeper layers reveal the evidence of aggregation of attention information over previous layers. The red boxes within the final attended images represent the ground truth bounding boxes for the query object annotated in the VG dataset. Each object may have multiple bounding boxes annotated by different annotators. The annotated answer is presented in the first column. The percentage for each method means the probability of the GT answer for corresponding method.

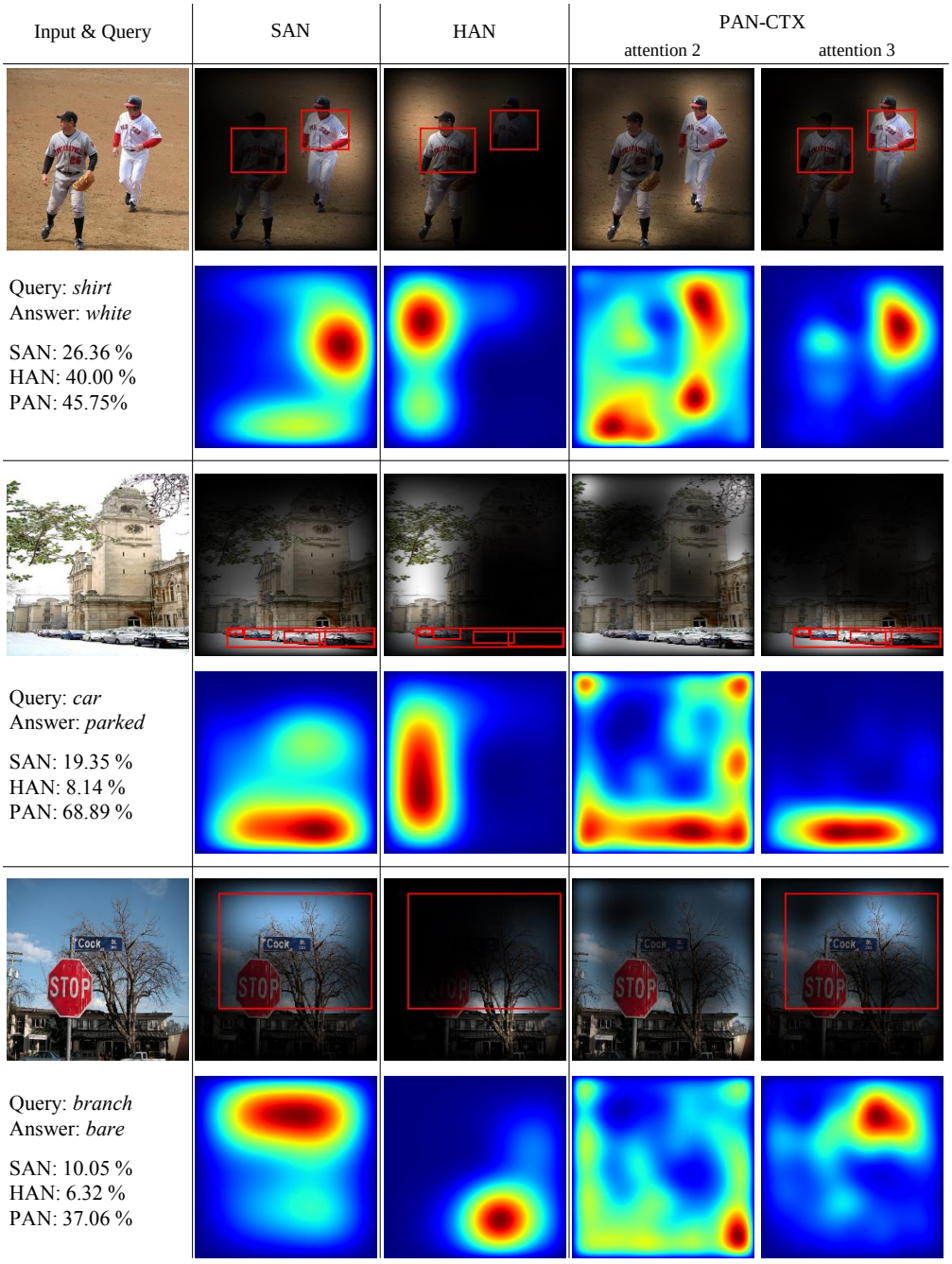


Figure 14: More qualitative results of SAN, HAN and PAN-CTX on the VG dataset.