
Hierarchical Attention Networks

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 network, which accurately attends to target objects of various scales and shapes in images through multiple stages. The proposed network enables multiple layers to estimate attention in a convolutional neural network (CNN). The hierarchical attention model gradually suppresses irrelevant regions in an input image using a progressive attentive process over multiple CNN layers. The attentive process in each layer determines whether to pass or suppress feature maps for use in the next convolution. We 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 various 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 [1, 2, 3, 4], image generation [5], image caption generation [6] and visual question answering [7, 8, 9].

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 [3].

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, which aggregates responses in a feature map weighted by their attention probabilities. This process results in a single attended feature vector. Refer to Section 3 for details of soft attention. Since the soft attention method is fully differentiable, the entire network can be trained end-to-end with standard backpropagation. However, it allocates attention probabilities to regions with a fixed size and shape as it maps from the receptive field of each location in the feature map. This makes the soft attention method inappropriate for real images, where objects involve significant variations in their scales, and shapes.

To overcome this limitation, we propose a novel attention network, referred to as **hierarchical attention network** (HAttNet), which has capability to precisely attend objects of different scales and

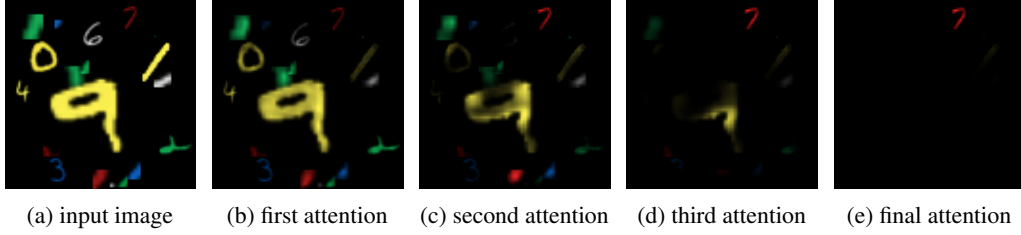


Figure 1: An example of reference problem (with the query 7 and the answer *red*). It shows that attention is gradually accumulated and forwarded to the subsequent layers in the network for resolving the reference problem. Distractors in the images are typically suppressed in earlier layers, which have smaller receptive fields and irrelevant objects are filtered out in the layers with larger receptive fields.

shapes by attaching attentive mechanisms within a convolutional neural network (CNN). In other words, the proposed network injects attention to intermediate feature maps and forwards the attended feature maps 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, attended regions are identified by augmenting attention layer by layer and eventually match the precise scale and shape of the target objects. The contribution of this work is three-fold:

- A new neural attention network that can gradually attend to areas of interest with accurate scale and shape through a hierarchical attentive mechanism
- A new idea of using local contexts to improve the stability of the hierarchical attention model
- Achievement of significant performance improvement over traditional soft attention approaches in query-specific attribute prediction tasks

The rest of this paper is organized as follows. We first review related work and the soft attention model in Section 2 and 3, respectively. In Section 4, we describe the proposed hierarchical attention network with local context information. We then present our experimental results on several datasets in Section 5 and conclude the paper in Section 6.

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 [6, 7, 8, 9, 10, 11, 12, 13]. For example, they have been used to handle sequences of variable lengths in neural machine translation models [10, 11] and manage memory access mechanisms for memory networks [12] and neural turing machines [13]. 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 [6, 7, 8, 9].

There have been several approaches iteratively performing attentive processes to resolve relations between targets. Yang *et al.* [7] 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, [12] and [13] incorporate attention mechanisms to memory cells iteratively to retrieve different values stored in the memory. Our proposed model is different in that attention information is predicted recursively from feature maps over multiple layers of CNN and accumulated layer by layer to capture the fine shapes of the target objects in the final attention map while the other approaches attend to the same feature map or memory.

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. [2] and [3] 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 [5] for image generation, where the network uses Gaussian filters for sampling. Spatial transformer networks (STNs) [1] 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. Because all these transformations used in [1] involve scale factors, STNs are capable of dealing with objects in different sizes. However, in Section 5, we show that STN is limited when there are multiple candidate regions for attention. We believe this is because the gradient signal gives only an indirect guide to the target object due to lack of supervision.

Training Attention Models The soft attention model sums up features weighted by the attention probabilities. The network becomes fully differentiable and thus trainable end-to-end by backpropagation. [6] and [14] 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, additional techniques, such as variational lower bound and Monte Carlo method resulting in the REINFORCE learning rule [15], were used to make networks trainable with a large search space defined by latent variables. Even though this hard attention model performed slightly better than soft attention on image caption generation, the size of the attended region is still fixed and even more restricted in shape than a soft attention model. Transformation based attention models [2, 3] are mostly trained by the REINFORCE learning rule because they make a hard decision of the attended region. STN [1] proposed a fully differentiable formulation and made it possible to train end-to-end. This formulation can also be applied to the other transformation based attention methods for facilitating training.

3 Soft Attention Model

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 $\mathbf{f} \in \mathbb{R}^{H \times W \times C}$ and a query q containing information of where to attend, a soft attention model first obtains an attended feature map $\hat{\mathbf{f}} \in \mathbb{R}^{H \times W \times C}$, where W is width, H is height, and C is the number of channels. The input feature map \mathbf{f} is generally a CNN output of an input image I , which is given by

$$\mathbf{f} = \text{CNN}(I). \quad (1)$$

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

$$s_{i,j} = g_{\text{att}}(f_{i,j}, q; \theta_{\text{att}}) \quad (2)$$

$$\alpha_{i,j} = \text{softmax}_{i,j}(\mathbf{s}), \quad 0 \leq \alpha_{i,j} \leq 1 \quad (3)$$

where $g_{\text{att}}(\cdot)$ is the attention network parameterized by θ_{att} and $\mathbf{s} = [s_{i,j}]$ is an attention score map. The attention score map is normalized with softmax to produce attention probabilities $\alpha_{i,j}$. Note that $g_{\text{att}}(\cdot)$ 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 $\hat{\mathbf{f}}$ at (i, j) . Then, the attended feature denoted by $f^{\text{att}} \in \mathbb{R}^C$ is computed by a weighted sum of features as

$$f^{\text{att}} = \sum_i^H \sum_j^W \hat{f}_{i,j} = \sum_i^H \sum_j^W \alpha_{i,j} f_{i,j}. \quad (4)$$

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.

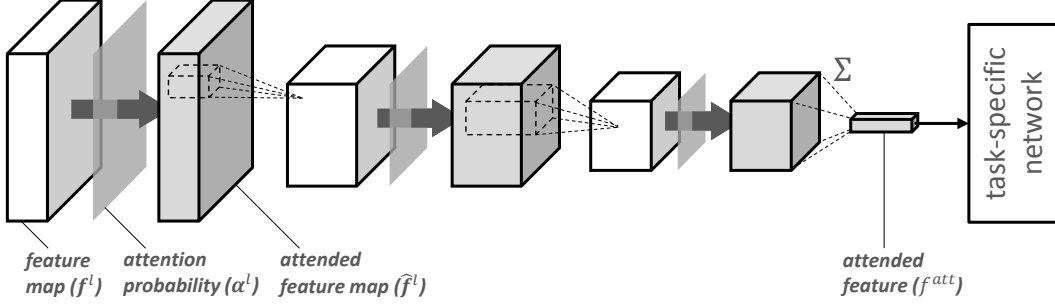


Figure 2: Overall architecture of the hierarchical attention network. Attentive processes are repeatedly applied to feature maps at multiple layers and the resulting attended feature maps are used as input feature maps for the next convolution layers in CNN.

4 Hierarchical Attention Network

To overcome the limitation of existing attention models in handling variable object scales and shapes, we propose a hierarchical attention mechanism. In the proposed model, irrelevant features at different scales are suppressed by multiple CNN layers and attention filtering steps, and computation is focused on the features corresponding to regions of interest. In this way, the network generates an attended feature map in each layer of CNN involving attentive process, and 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 recursive process allows us to estimate precise details of attention areas while maintaining deep representations appropriate for high-level inference tasks.

4.1 Recursive Attentive Process

Let $\mathbf{f}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ 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 $f_{i,j}^l \in \mathbb{R}^{C_l}$ be a feature at (i, j) of the feature map \mathbf{f}^l . In the proposed hierarchical attention network, an attentive process is applied to multiple layers of CNN and we obtain the attended feature map $\hat{\mathbf{f}}^l = [\hat{f}_{i,j}^l]$, which is given by

$$\hat{f}_{i,j}^l = \alpha_{i,j}^l f_{i,j}^l. \quad (5)$$

Here, the attention probability $\alpha^l = [\alpha_{i,j}^l]$ for a feature $f_{i,j}^l$ is calculated by

$$s_{i,j}^l = g_{\text{att}}^l(f_{i,j}^l, q; \theta_{\text{att}}^l) \quad \text{and} \quad \alpha_{i,j}^l = \begin{cases} \text{softmax}_{i,j}(s^l) & \text{if } l = L \\ \sigma(s_{i,j}^l) & \text{otherwise} \end{cases}, \quad (6)$$

where $g_{\text{att}}^l(\cdot)$ denotes the attention function with a set of parameters θ_{att}^l for layer l , $s_{i,j}^l$ is the attention score at (i, j) in layer l , and $\sigma(\cdot)$ is a sigmoid function. An attention probability at each location is independent of others in the same feature map, where a sigmoid function is typically employed to make attention probabilities set as a value between 0 and 1. For the last layer of attention, we use a softmax function over the entire spatial region to attend small regions of interest only.

Unlike the soft attention model, the attended feature map $\hat{\mathbf{f}}^l$ 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

$$\mathbf{f}^{l+1} = g_{\text{CNN}}^{l+1}(\hat{\mathbf{f}}^l; \theta_{\text{CNN}}^{l+1}) = g_{\text{CNN}}^{l+1}(\mathbf{r}(\alpha^l, C^l) \odot \mathbf{f}^l; \theta_{\text{CNN}}^{l+1}) \quad (7)$$

where $g_{\text{CNN}}^l(\cdot)$ is an operation at layer l in CNN parameterized by θ_{CNN}^l , $\mathbf{r}(\alpha^l, C^l)$ denotes replication of α^l C_l times in its channel direction, and \odot denotes elementwise multiplication.

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

$$\mathbf{f}^{\text{att}} = \sum_i^H \sum_j^W \hat{f}_{i,j}^L. \quad (8)$$

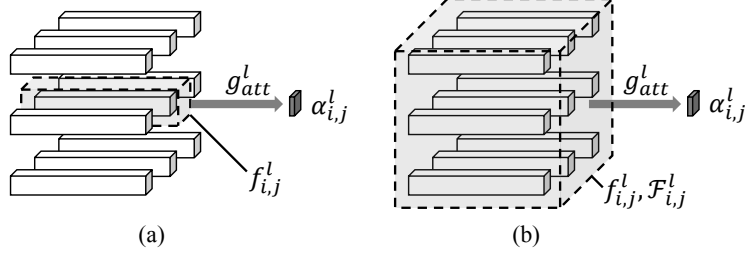


Figure 3: Attention estimation (a) without local context and (b) with local context. In (a), $\alpha_{i,j}^l$ is predicted from $f_{i,j}^l$ only while its spatially adjacent features are also used to estimate $\alpha_{i,j}^l$ in (b).

The attended feature f^{att} obtained by such process is then used as the input to the task specific network such as classification network as illustrated in Figure 2.

4.2 Multi-Resolution Attention Estimation

In Eq. (7), the resolution of attention probability map α^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 α^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.

Our attentive process is opposite to more intuitive coarse-to-fine approaches. However, what happens in our hierarchical attention network is similar in some sense to the coarse-to-fine methods because it learns to accumulate multi-resolution attention probability maps to identify the final attended regions. It forwards the high-resolution features corresponding to candidate regions of interest based on the observation of low-level features at the beginning, and then aggregates attention to the relevant regions of low-resolution (equivalently, filters out the irrelevant regions from low-resolution layer) based on high-level understanding later. We claim that a properly trained hierarchical attention network can estimate fine-grained attention probability maps through the multi-resolution attentive processing. To incorporate local context information in each layer, we observe neighborhood features to estimate attention probability map, which is critical for the effectiveness of the hierarchical attention process and facilitate the training procedure. We will elaborate this technique in the next subsection. An additional benefit of our hierarchical attention network is that it is more straightforward for inference since it is a pure feedforward network.

4.3 Local Context

The naïve implementation of our hierarchical attention network predicts an attention probability $\alpha_{i,j}^l$ based solely on the feature $f_{i,j}^l$ at a single location. However, the attended feature $\hat{f}_{i,j}^l$ is supposed to refer to its neighborhood through convolution or pooling operations in the subsequent layers. This creates a gap between the attention process and the subsequent operations in CNN in terms of their capability of observable regions. This gap could make the hierarchical attention network fail to identify regions of interest accurately and reduces the effectiveness in training our attention network.

We bridge this gap by allowing the attention network at each layer to observe a local context of the target feature during attention prediction. The local context $\mathcal{F}_{i,j}^l$ of a feature $f_{i,j}^l$ is composed of its spatially adjacent features. For example, the local context can be given by $\mathcal{F}_{i,j}^l = \{f_{s,t}^l | i - \delta \leq s \leq i + \delta, j - \delta \leq t \leq j + \delta\}$ as illustrated in Figure 3. The attention score is now predicted by the attention network with local context as

$$s_{i,j}^l = g_{\text{att}}^l(f_{i,j}^l, \mathcal{F}_{i,j}^l, q; \theta_{\text{att}}^l). \quad (9)$$

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.

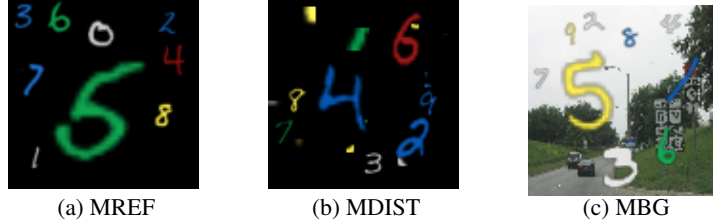


Figure 4: Example of the MREF datasets.

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

(a) Color prediction accuracy [%]				(b) True-positive strength [%]			
	MREF	MDIST	MBG		MREF	MDIST	MBG
STN	39.10	38.32	32.27	Uniform	2.34	2.35	2.39
SOFT	82.94	75.73	53.77	SOFT	13.61	12.56	6.73
HARD	81.84	78.49	55.84	HARD	13.95	13.81	7.64
HAttNet	95.92	91.65	69.46	HAttNet	17.39	13.10	8.62
HAttNet-CTX	98.51	96.02	81.01	HAttNet-CTX	22.59	22.80	11.01

4.4 Training Hierarchical Attention Networks

Training a hierarchical attention network is as simple as training a soft attention network because every operation within the network is differentiable as in the soft attention. The entire network is trained end-to-end by the standard backpropagation based on a task-specific loss.

When we train the proposed hierarchical attention network 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. On the other hand, fine-tuning CNN is not required for the soft attention although it generally improves performance.

5 Experiments

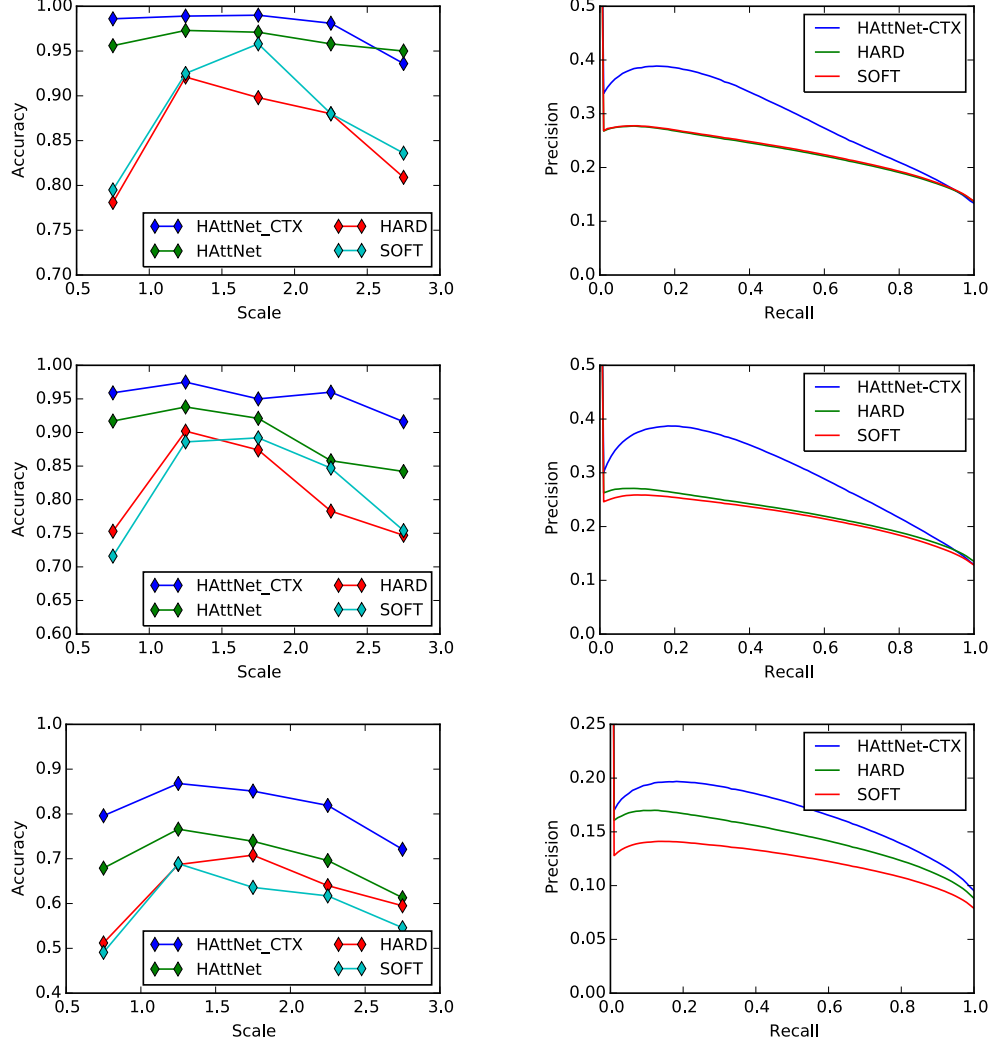
5.1 MNIST Reference

Datasets We conduct experiments on a synthetic dataset created from MNIST [16]. 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 {green, yellow, white, red, 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 changing 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 [17].

Experimental Settings We implement the proposed network with and without the local context observation referred to as HAttNet-CTX and HAttNet, respectively. In addition, three different baselines, soft attention (SOFT) [6], hard attention (HARD) [6], spatial transformer network (STN) [1], are used for comparisons. The implementations of SOFT and STN follow the descriptions in [6] and [1], respectively, while HARD is trained by optimizing the marginal log-likelihood loss since it should be more accurate and feasible due to small search space.

The architecture of image encoding network in SOFT and HARD and localization network in STN are all identical. 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). We used a single fc layer for classification because the task requires simple color prediction. Every model is trained from scratch.

Results Table 1a presents color prediction accuracy of all compared algorithms. It is obvious that HAttNet outperforms all the previous approaches with significant margins and HAttNet-CTX further improves the performance by exploiting the local contexts for attention estimation.



(a) The accuracies of the four best models on the test subsets in different scales.

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

Figure 5: Analysis of algorithms on MREF (top), MDIST (middle), and MBG (bottom).

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 four best performing models. The results are presented in Figure 5a, where SOFT and HARD 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. In contrast, HAttNet and HAttNet-CTX are robust to scale variations due to their multi-scale attention mechanism especially when the local contexts are incorporated.

We also evaluate the quality of attention maps using two complementary criteria: true-positive strength and precision-recall (PR) curve. The former estimates how strong attention is given to proper location by computing the sum of the attention probabilities within the ground-truth segmentation divided by the sum of all attention probabilities. The latter measures the overlaps between ground-truth segmentations and binarized segmentation predictions constructed with different thresholds. The true-positive strengths of all algorithms are presented in Table 1b and the PR-curves are illustrated in Figure 5b. Note that the proposed model with the local context observation gives the best results with noticeable gaps to the other methods in terms of both criteria. These results indirectly suggest that HAttNet-CTX construct more accurate shapes of attended regions than all other attention models.

Table 2: The results of the attribute prediction measured in weighted mAP [%].

	attention only	w/ prior
SOFT	24.84	28.88
HARD	23.81	28.04
HAttNet-CTX	27.04	30.38

The qualitative results of the proposed algorithm and two baselines are presented in Figure 6 (MREF and MDIST) and Figure 7 (MBG). The proposed model yields accurate attention regions eventually by gradually augmenting attention and suppressing irrelevant regions in the image. In contrast, the baseline models attend to the target objects only once at the top layer resulting in coarse attention in size and shape.

5.2 Attribute Prediction on Visual Genome

Dataset Visual Genome (VG) [18] 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 (*e.g.* an object category), we predict the binary attributes of individual objects specified by a query. We used 488 object classes and 274 attribute classes that appear more than 100 times. A total of 45,393 images with 255,118 object attribute labels are used for our experiment, and they are split into training, validation and test sets each containing 25,393, 10,000 and 10,000 images.

The dataset is challenging for several reasons. First, the annotations are noisy and incomplete because the object and attribute annotations are only extracted from the annotated captions by crowdsourcing. There are many missing objects and attributes not mentioned in any captions. Second, scales of objects largely vary and the attributes may be associated with very small objects. Third, the attribute labels for objects are highly skewed. For each object, only about 1.2 attributes out of 274 are annotated as positive on average while all others are negative. Finally, some of the annotated objects and attributes are even invisible or indistinguishable using the object query (*e.g.* *air*, *it* and *something*).

Experimental Settings and Results We compared our algorithm with HARD and SOFT. All the networks share the same CNN architecture of VGG-16 layer net [19], which is pretrained on ImageNet [20] and is further fine-tuned on the VG dataset for the attribute prediction. For HARD and SOFT, an attention layer is attached to the last pooling layer in VGG-16 while HAttNet stacks an additional attention layer on top of each of the last three pooling layers in VGG-16.

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 [21]. 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. Figure 8 presents the qualitative results of the proposed network and SOFT on VG dataset.

6 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

- [1] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS. (2015) 2008–2016
- [2] Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. In: ICLR. (2015)

- [3] Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: NIPS. (2014) 2204–2212
- [4] Larochelle, H., Hinton, G.E.: Learning to combine foveal glimpses with a third-order boltzmann machine. In: NIPS. (2010) 1243–1251
- [5] Gregor, K., Danihelka, I., Graves, A., Wierstra, D.: Draw: A recurrent neural network for image generation. In: ICML. (2015) 1462–1471
- [6] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. (2015)
- [7] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. arXiv preprint arXiv:1511.02274 (2015)
- [8] Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Deep compositional question answering with neural module networks. In: CVPR. (2016)
- [9] Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv preprint arXiv:1511.05234 (2015)
- [10] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR. (2015)
- [11] Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: EMNLP. (2015)
- [12] Weston, J., Chopra, S., Bordes, A.: Memory networks. In: ICLR. (2015)
- [13] Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
- [14] Zaremba, W., Sutskever, I.: Reinforcement learning neural turing machines. arXiv preprint arXiv:1505.00521 (2015)
- [15] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**(3-4) (1992) 229–256
- [16] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11) (1998) 2278–2324
- [17] Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision* (2014) 1–20
- [18] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332 (2016)
- [19] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *ICLR* (2015)
- [20] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
- [21] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International journal of computer vision* **88**(2) (2010) 303–338

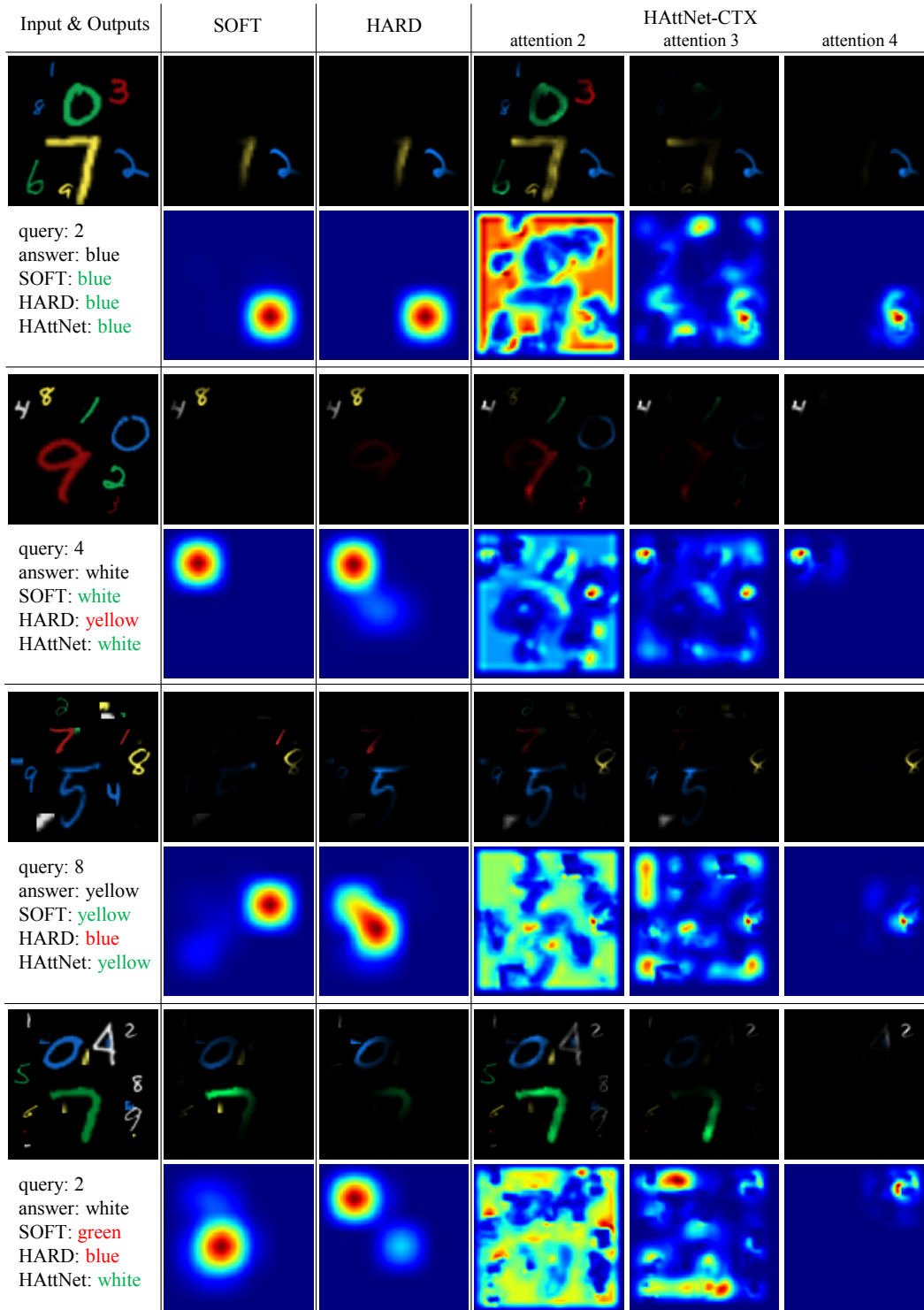


Figure 6: The qualitative results of SOFT, HARD and HAttNet-CTX on the MREF and MDIST datasets. For each example, attended images are shown in the first row and the corresponding attention maps are shown on the second row. In case of the hierarchical attention network, the last three attention maps are visualized. As can be seen, attention map at deeper layers reveal the evidence of aggregation over earlier attention maps.


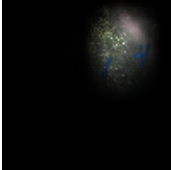
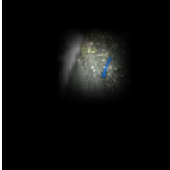





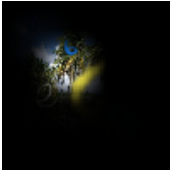








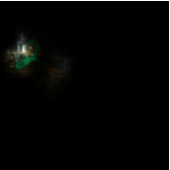





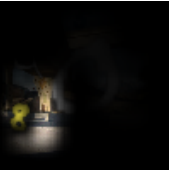
Input & Outputs	SOFT	HARD	HAttNet-CTX		
			attention 2	attention 3	attention 4
 <p>query: 1 answer: blue SOFT: blue HARD: blue HAttNet: blue</p>					
 <p>query: 6 answer: blue SOFT: yellow HARD: blue HAttNet: blue</p>					
 <p>query: 2 answer: green SOFT: blue HARD: blue HAttNet: green</p>					
 <p>query: 8 answer: yellow SOFT: white HARD: white HAttNet: yellow</p>					

Figure 7: The qualitative results of SOFT, HARD and HAttNet-CTS on the MBG dataset. 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 hierarchical attention network, the last three attention maps are visualized. As can be seen, attention map at deeper layers reveal the evidence of aggregation over earlier attention maps.






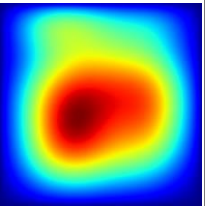
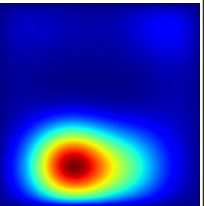
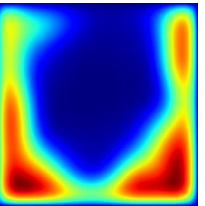
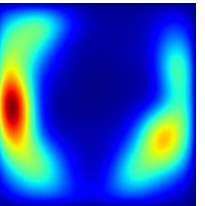





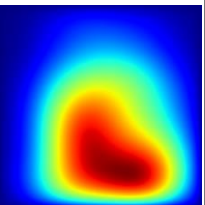
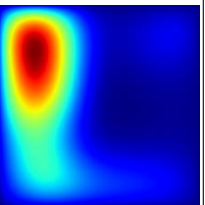
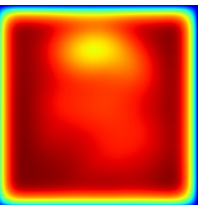
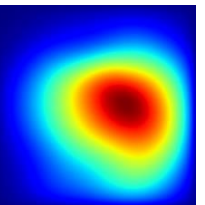



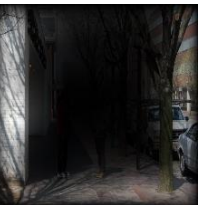

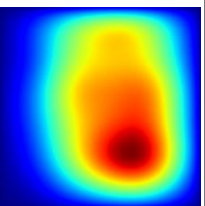
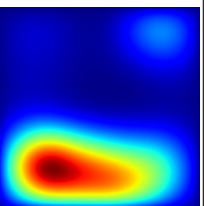
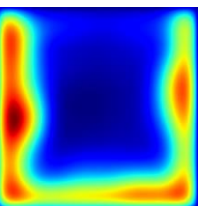
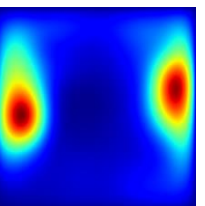
Input & Query	SOFT	HARD	HAttNet-CTX	
			attention 2	attention 3
				
Query: bush				
				
Query: sweater				
				
Query: wall				

Figure 8: The qualitative results of SOFT, HARD and HAttNet-CTX on the VG dataset. For each example, the attended images are presented in the first row while their attention maps are shown in the second row. In the case of the hierarchical attention network, 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.