SAM: Pushing the Limits of Saliency Prediction Models

Marcella Cornia[1] Lorenzo Baraldi[1] Giuseppe Serra[2] Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia [2]University of Udine
{marcella.cornia,lorenzo.baraldi,rita.cucchiara}@unimore.it giuseppe.serra@uniud.it

Abstract

The prediction of human eye fixations has been recently gaining a lot of attention thanks to the improvements shown by deep architectures. In our work, we go beyond classical feed-forward networks to predict saliency maps and propose a Saliency Attentive Model which incorporates neural attention mechanisms to iteratively refine predictions. Experiments demonstrate that the proposed strategy overcomes by a considerable margin the state of the art on the largest dataset available for saliency prediction. Here, we provide experimental results on other popular saliency datasets to confirm the effectiveness and the generalization capabilities of our model, which enable us to reach the state of the art on all considered datasets.

1. Saliency Attentive Model (SAM)

In the last decades, a significant research effort has been dedicated to the development of saliency prediction models, which can predict human eye fixations. It has been shown that emulating where humans look in a scene can enhance many vision-based applications, ranging from image captioning [5, 6] to automatic cropping [7]. With the advent of deep learning, saliency prediction has achieved a strong improvement, thanks to both novel architectures and large-scale datasets [11, 3, 13]. Even though these approaches overcame by a big margin hand-crafted methods, the use of machine attention models has been rarely investigated in this task. We recently proposed a Saliency Attentive Model (SAM) [4] which, in contrast, incorporates attentive mechanisms to iteratively refine saliency predictions. Overall, it is composed by three main components: a Dilated Convolutional Network that extracts feature maps from the input image, an Attentive Convolutional LSTM which recurrently enhances saliency features and a learned prior module that incorporates the human-gaze center bias in the final predictions. The overall architecture is shown in Fig. 1.

Dilated Convolutional Network. Deep saliency architectures are usually built over a pre-trained CNN that extracts feature maps from input images. One of the main drawbacks of this approach is that it considerably rescales the input image, thus worsening the saliency prediction performance. To limit this rescaling effect, we use a Dilated CNN that, thanks to dilated convolutions and modifications of standard CNN architectures, produces saliency maps with an increased output size. In particular, we propose two different variations of our model: one based on VGG-16, and the other based on ResNet-50. Thanks to this strategy, the predicted saliency maps are rescaled, for both versions, by a factor of 8 instead of 32 as in the original CNNs.

Attentive Convolutional LSTM. The feature maps coming from the dilated network are then input to an Attentive Convolutional model, which recurrently process saliency features at different locations. We extend the traditional LSTM to work on spatial features by substituting dot products with convolutional operations, so that hidden states are feature stacks instead of vectors. Moreover, we exploit the sequential nature of LSTM to process features in an iterative way. The input of the LSTM is computed, at each step, through an attentive mechanism which focuses on different regions of the image. An attention map is generated by convolving the previous hidden state and the input (*i.e.* a stack of feature maps); once normalized through the *softmax* operator, this is applied to the input with an element-wise product. The result of this operation is a refined stack of features which is iteratively fed to the LSTM. After a fixed number of iterations, the last hidden state is taken as the output of this module.

Learned Priors. Finally, the output of the Attentive LSTM is combined with multiple learned priors which are used to model the center bias present in the human-eye fixations. Differently from existing works, which included predefined priors, we let the network learn its own priors. To reduce the number of parameters and facilitate the learning, we constraint that each prior should be a 2d Gaussian function, whose mean and covariance matrix are freely learnable. In this manner, priors are inferred purely from data, without relying on assumptions from biological studies.

Loss Function. During the training phase, the network is encouraged to minimize a combination of different cost functions, taking into account different quality aspects that

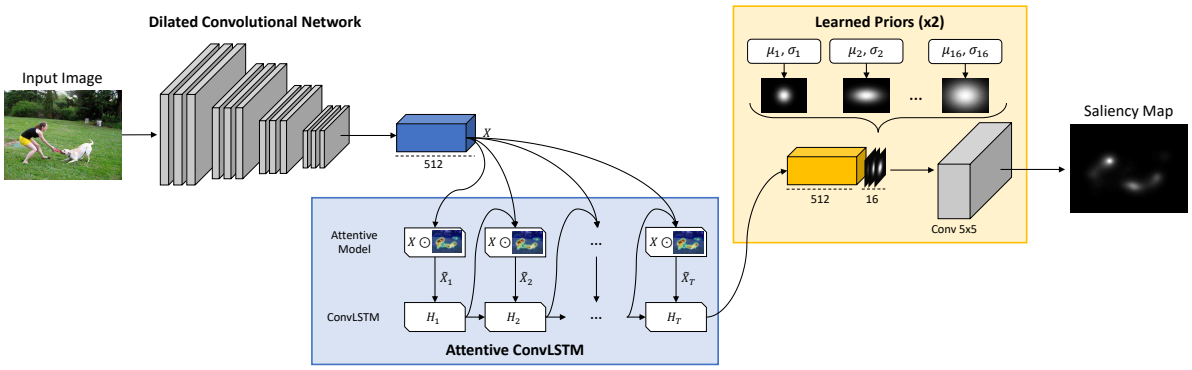

Figure 1. Overview of our Saliency Attentive Model (SAM).

predictions should meet. In particular, our loss function is given by a linear combination of three saliency evaluation metrics: the Normalized Scanpath Saliency (NSS), the Linear Correlation Coefficient (CC) and the Kullback-Leibler Divergence (KL-Div), all commonly used to evaluate saliency prediction models. We refer the reader to [2] for an extensive analysis of these saliency metrics and those used in the experimental section.

2. Experimental Results

Several saliency prediction datasets are currently available in literature. The largest one is SALICON [11] composed by $20,000$ images with corresponding saliency maps computed from mouse movements. Recently, a new version of this dataset has been released in which authors replace the original velocity-based fixation detection algorithm, resulting in more eye-like fixations. Here, we extend the work in [4] by using both versions of this dataset and comparing the results of our model trained on the two annotations. Furthermore, we also investigate on several other datasets.

Tables 1 and 3 report the results of both versions of our model (SAM-VGG and SAM-ResNet) on the two releases of the SALICON dataset, respectively. As it can be seen, our model overcomes all existing methods on both versions of SALICON and, as expected, the ResNet version obtains better results than the VGG-based model. Nevertheless, the version based on VGG-16 is still able to surpass the competitors on almost all the considered metrics. Fig. 2 shows some qualitative results on sample images from the SALICON dataset and visually highlights the differences between the two versions of the considered dataset.

Starting from our model trained on the two releases of the SALICON, we also evaluate the effectiveness of the proposal on other four popular saliency datasets: MIT1003 [12], TORONTO [1], PASCAL-S [14] and DUT-OMRON [20]. For a fair comparison with other methods, we do not finetune our model on a subset of these datasets. The comparison results are reported in Table 2. Again, we

	CC	sAUC	AUC	NSS
SAM-Resnet	**0.842**	**0.779**	0.883	**3.204**
SAM-VGG	0.825	0.774	0.881	3.143
ML-Net [3]	0.743	0.768	0.866	2.789
SalGAN [16]	0.781	0.772	0.781	2.459
SalNet [17]	0.622	0.724	0.858	1.859
DeepGazeII [13]	0.509	0.761	**0.885**	1.336

Table 1. Comparison results on SALICON 2015 test set [11]. Methods are sorted by the NSS metric.

observe that our model is able to quantitatively overcome the drawbacks of different existing proposals. As a side note, here the performance of the VGG-based model is often very similar to that of the ResNet-based one. Also, it shall be observed that the 2017 version of SALICON shows better generalization capabilities on all metrics except from NSS. This can be partially explained by the fact that the ground-truth maps of SALICON 2015 are less blurred than in the second version of the dataset: this helps the NSS measure, which normalizes the prediction to have zero mean and unit variance, thus increasing the weight of predicted pixels when the prediction is less blurred.

3. Conclusion

We gave a short overview of our Saliency Attentive Model (SAM). This incorporates dilated convolutions, an attentive mechanism and learned prior maps: the combination of these components has shown to overcome the state of the art on saliency prediction on different datasets, thus confirming the effectiveness of the approach.

References

[1] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *ANIPS*, pages 155–162, 2006.

[2] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency

	MIT1003				DUT-OMRON				TORONTO				PASCAL-S			
	CC	Sim	AUC	NSS	CC	Sim	AUC	NSS	CC	Sim	AUC	NSS	CC	Sim	AUC	NSS
Itti [10]	0.33	0.32	0.77	1.10	0.46	0.39	0.83	1.54	0.48	0.45	0.80	1.30	0.42	0.36	0.82	1.30
GBVS [8]	0.42	0.36	0.83	1.38	0.53	0.43	0.87	1.71	0.57	0.49	0.83	1.52	0.45	0.36	0.84	1.36
eDN [18]	0.41	0.30	0.85	1.29	-	-	-	1.33	0.50	0.40	0.85	1.25	-	-	-	1.42
Mr-CNN [15]	0.38	0.35	0.80	1.36	-	-	-	-	0.49	0.47	0.80	1.41	-	-	-	-
DVA [19]	0.64	0.50	0.87	2.38	0.67	0.53	0.91	3.09	0.72	0.58	**0.86**	2.12	0.66	0.52	0.89	2.26
SAM-VGG$_{2015}$	0.61	0.52	0.88	2.25	0.65	0.53	0.91	2.91	0.69	0.59	**0.86**	2.14	0.72	0.60	**0.90**	**2.48**
SAM-VGG$_{2017}$	0.65	0.52	**0.89**	2.33	0.69	0.53	0.91	2.95	**0.74**	**0.63**	**0.86**	**2.15**	0.73	**0.61**	0.89	2.31
SAM-ResNet$_{2015}$	0.65	**0.54**	0.88	**2.48**	0.69	**0.56**	0.91	**3.21**	0.69	0.59	**0.86**	2.12	0.69	0.59	0.89	2.34
SAM-ResNet$_{2017}$	**0.66**	0.53	**0.89**	2.35	**0.70**	0.54	**0.92**	2.97	**0.74**	0.62	**0.86**	2.14	**0.74**	**0.61**	**0.90**	2.34

Table 2. Comparison results on MIT1003 [12], DUT-OMRON [20], TORONTO [1] and PASCAL-S [14] dataset. Results of comparison methods are from [19].

	CC	Sim	AUC	NSS
SAM-ResNet	**0.899**	**0.793**	**0.865**	**1.990**
SAM-VGG	0.891	0.786	0.864	1.971
EAD [9]	0.871	0.760	0.852	1.896
SalGAN [16]	0.844	0.728	0.857	1.816
SalNet [17]	0.763	0.639	0.840	1.555

Table 3. Comparison results on SALICON 2017 test set [11]. Methods are sorted by the NSS metric. Results of comparison methods are from [9].

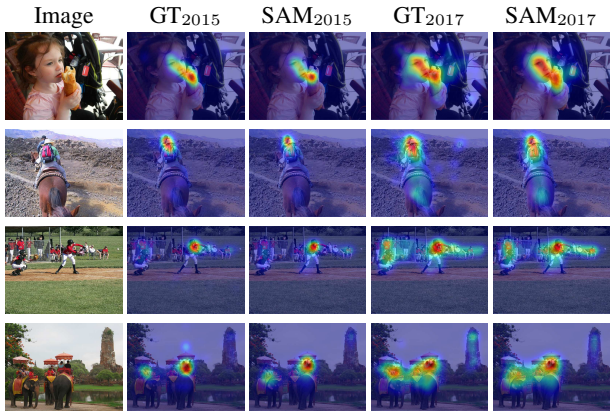

Figure 2. Qualitative results on both 2015 and 2017 releases of the SALICON dataset [11].

models? *arXiv preprint arXiv:1604.03605*, 2016.

[3] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Multi-level Net: A Visual Saliency Prediction Model. In *ECCV Workshops*, 2016.

[4] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*, 2017.

[5] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Visual Saliency for Image Captioning in New Multimedia Services. In *ICME Workshops*, 2017.

[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. *ACM Trans. Multimedia Comput. Commun. and Appl.*, 14, 2018.

[7] M. Cornia, S. Pini, L. Baraldi, and R. Cucchiara. Automatic image cropping and selection using saliency: An application to historical manuscripts. In *Digital Libraries and Multimedia Archives*, volume 806, 2018.

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *ANIPS*, 2006.

[9] S. He, N. Pugeault, Y. Mi, and A. Borji. What Catches the Eye? Visualizing and Understanding Deep Saliency Models. *arXiv preprint arXiv:1803.05753*, 2018.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998.

[11] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *CVPR*, 2015.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[13] M. Kümmerer, T. S. Wallis, and M. Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

[14] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[15] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, 2015.

[16] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. In *CVPR Workshops*, 2017.

[17] J. Pan, K. McGuinness, E. Sayrol, N. O'Connor, and X. Giró-i Nieto. Shallow and Deep Convolutional Networks for Saliency Prediction. In *CVPR*, 2016.

[18] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014.

[19] W. Wang and J. Shen. Deep visual attention prediction. *IEEE Trans. Image Process.*, 27(5):2368–2378, 2018.

[20] C. Yang, L. Zhang, R. X. Lu, Huchuan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.