PREDICTING HUMAN EYE FIXATIONS VIA AN LSTM-BASED SALIENCY ATTENTIVE MODEL

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra and Rita Cucchiara

University of Modena and Reggio Emilia, Italy - name.surname@unimore.it

ICVSS 2017 — Sicily ~ 09-15 July — International Computer Vision Summer School

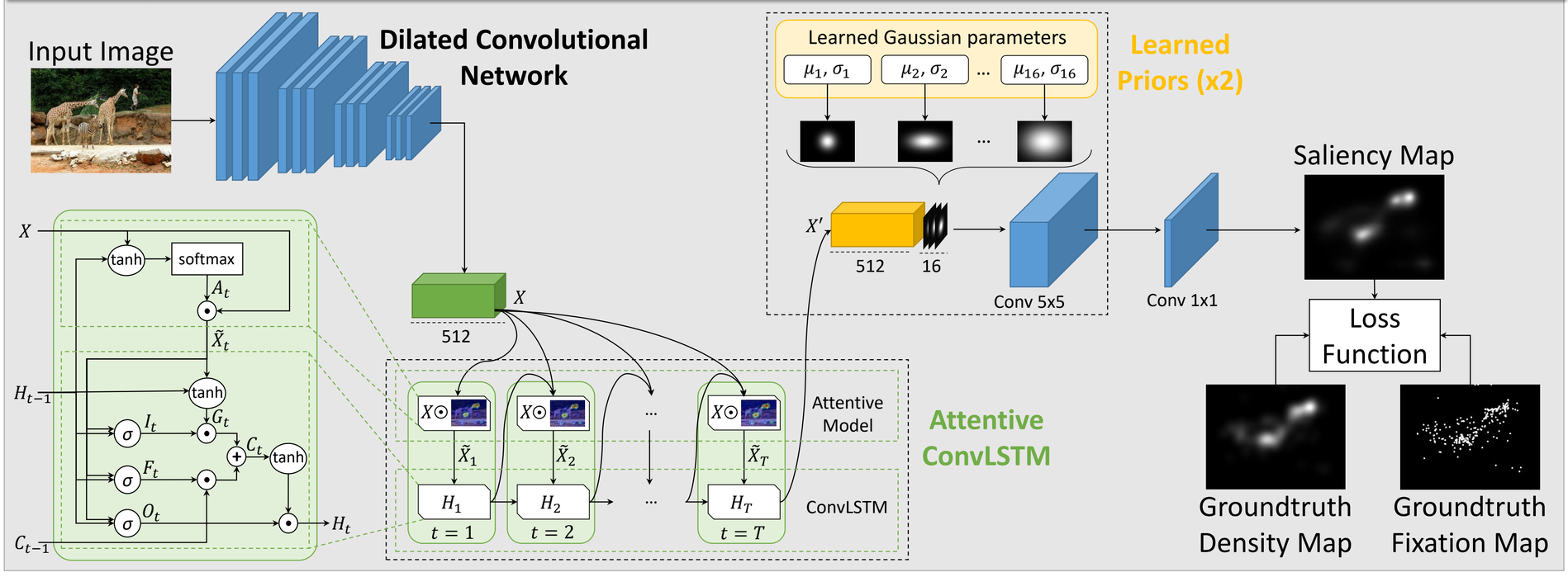
Source code available at:

Abstract

In this work we go beyond standard approaches to saliency prediction and we present a novel model which can predict accurate saliency maps by incorporating neural attentive mechanisms.

The core of our solution is a Convolutional LSTM that focuses on the most salient regions of the input image to iteratively refine the predicted saliency map. Additionally, to tackle the center bias present in human eye fixations, our model can learn a set of prior maps generated with Gaussian functions.

Saliency Attentive Model (SAM)



SAM Architecture

Attentive ConvLSTM

- Extension of the traditional LSTM to work on spatial features by substituting dot products with convolutional operations.
- Exploitation of the sequential nature of the LSTM to process features in an iterative way, without the concept of time.

The input of the LSTM layer $\tilde{X}_t$ is computed through an **attentive mechanism** which produces an attention map from the previous hidden state H_{t-1} of the LSTM and the input X

$$Z_t = V_a * \tanh(W_a * X + U_a * H_{t-1} + b_a)$$

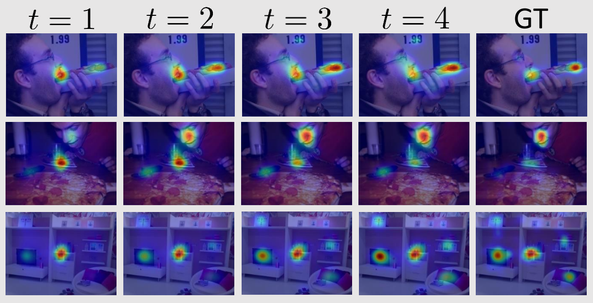
The output of this operation is a 2-d map from which we compute a normalized spatial attention map through the *softmax* operator

$$A_t^{ij} = p(att_{ij}|X, H_{t-1}) = \frac{\exp(Z_t^{ij})}{\sum_i \sum_j \exp(Z_t^{ij})}$$

where A_t^{ij} is the element of the attention map in position (i, j).

The attention map is applied to the input with an element-wise product between each channel of the feature maps and the attention map

$$\tilde{X}_t = A_t \odot X.$$

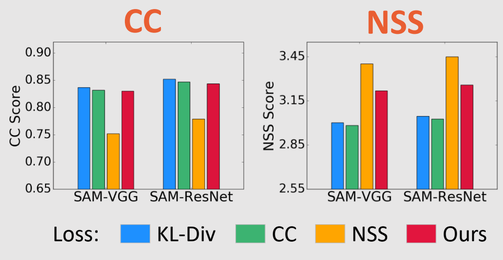


Learned Priors

- Our network is able to learn the center bias present in eye fixations, without the need to integrate this information manually.
- The model learns means and variances of a set of Gaussian functions with diagonal covariance matrix and produces a prior map for each function.

Loss Function

To take different quality aspects into account, we define a new loss function given by a linear combination of three saliency evaluation metrics: the **NSS**, the **CC** and the **KL-Div**.



Dilated Convolutional Network

We build two different versions of our model, one based on the **VGG-16** and the other based on the **ResNet-50**. To limit the rescaling, we employ dilated convolutions thus obtaining saliency maps rescaled by a factor of 8 instead of 32.

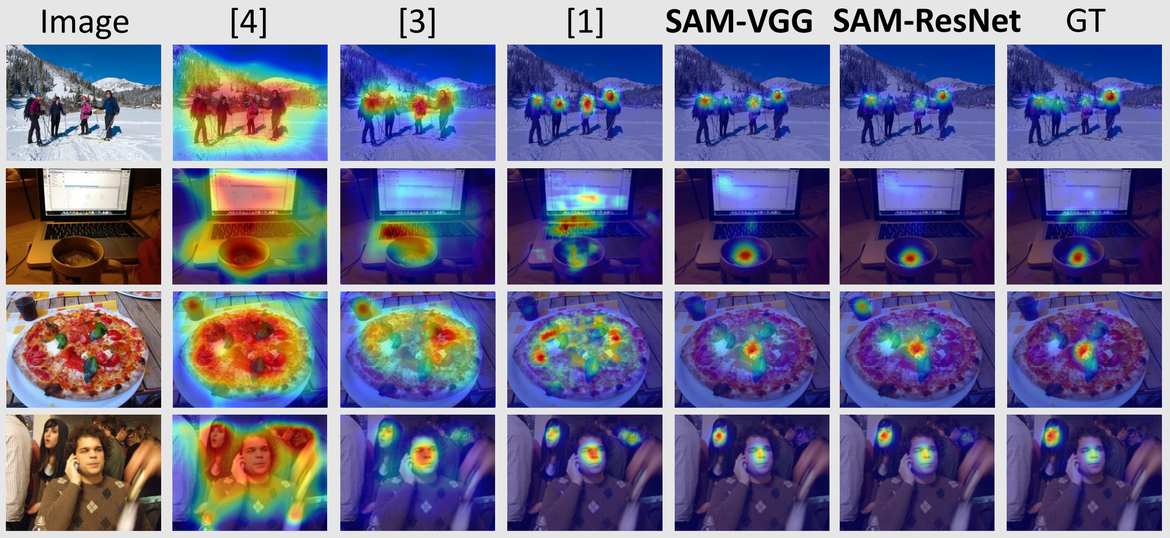
Experimental results

Contribution of the Attentive ConvLSTM

	T	CC	sAUC	AUC	NSS
SAM-VGG	1	0.821	0.777	**0.884**	3.168
SAM-VGG	2	0.827	0.777	0.883	3.224
SAM-VGG	3	0.828	0.781	0.883	**3.226**
SAM-VGG	4	**0.830**	**0.782**	0.883	3.219
SAM-ResNet	1	0.785	0.737	0.879	3.050
SAM-ResNet	2	0.829	0.764	**0.886**	3.214
SAM-ResNet	3	0.842	0.779	**0.886**	3.256
SAM-ResNet	4	**0.844**	**0.787**	**0.886**	**3.260**

Overall performance when using the output of the Attentive ConvLSTM at different timestep as input for the rest of the model.

The refinement performed by the attentive model results in better performance.



State-of-the-art results on three different datasets, overcoming all other methods by a big margin especially on SALICON, the biggest dataset available for saliency.

Results on SALICON Dataset

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.84**	**0.78**	0.88	**3.20**
SAM-VGG	0.83	0.77	0.88	3.14
ML-Net [1]	0.74	0.77	0.87	2.79
SU [2]	0.78	0.76	0.88	2.61
SalNet [3]	0.62	0.72	0.86	1.86
DeepGazeII [4]	0.51	0.76	**0.89**	1.34

Results on MIT300 Dataset

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.78**	0.70	0.87	**2.34**
SAM-VGG	0.77	**0.71**	0.87	2.30
DeepFix [5]	**0.78**	**0.71**	0.87	2.26
ML-Net [1]	0.67	0.70	0.85	2.05
SalNet [3]	0.58	0.69	0.83	1.51
DeepGazeII [4]	0.52	0.72	**0.88**	1.29

Results on CAT2000 Dataset

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.89**	**0.58**	**0.88**	**2.38**
SAM-VGG	**0.89**	**0.58**	**0.88**	**2.38**
DeepFix [5]	0.87	**0.58**	0.87	2.28

References

[1] Cornia, et al. "A Deep Multi-Level Network for Saliency Prediction." *ICPR*, 2016.
[2] Kruthiventi, et al. "Saliency Unified: A deep architecture for eye fixation prediction and salient object segmentation." *CVPR*, 2016.
[3] Pan, et al. "Shallow and Deep Convolutional Networks for Saliency Prediction." *CVPR*, 2016.
[4] Kümmerer, et al. "DeepGaze II: Reading fixations from deep features trained on object recognition." *arXiv:1610.01563*, 2016.
[5] Kruthiventi, et al. "DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations." *arXiv:1510.02927*, 2015.