A Deep Multi-Level Network for Saliency Prediction

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra and Rita Cucchiara

University of Modena and Reggio Emilia

Problem Statement

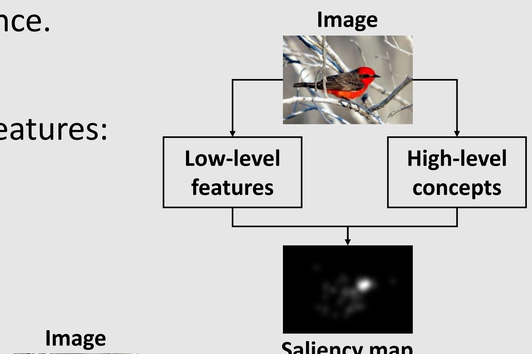
Classical algorithms for saliency prediction focused on identifying fixation points that human viewer would focus on at first glance.

Conventional Saliency

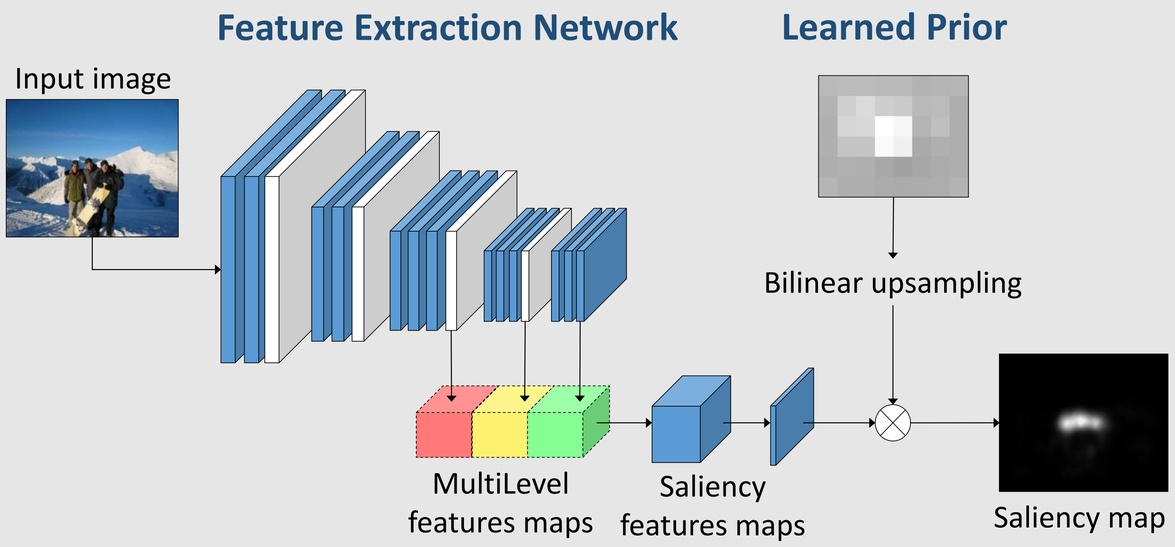
- Extraction of hand-crafted and multi-scale features:
  - Lower-level features
  - Higher-level concepts
    - faces, people, text, horizon, etc.
- Difficult to combine all these factors.

Deep Saliency

- Fully Convolutional networks directly predict saliency maps given by a non-linear combination of high level feature maps extracted from the last convolutional layer.



Proposed Architecture



Feature Extraction and Encoding Network

- We build our architecture on the popular VGG-16 model.
- To limit rescaling, the last pooling stage is removed and the stride of the last but one pooling layer is decreased.
- We take feature maps at three different locations of the FCN, and concatenate them to form a single tensor.
- A 3 x 3 convolutional layer learns 64 saliency-specific feature maps, then a 1 x 1 convolution learns to weight each map to produce a temporary saliency prediction.

Learned Prior

- Instead of using pre-defined priors as in previous works, we let the network learn its own custom prior.
- A coarse mask, which has a much smaller size of the saliency map, is learned.
- Then it is upsampled and applied to the predicted saliency map with pixel-wise multiplication.

Loss Function

Three objectives:

- Predictions should be pixel-wise similar to ground truth.
- Predicted maps should be invariant to their maximum.
- The loss should give the same importance to high and low GT values.

$$L(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left\| \frac{\frac{\phi(\mathbf{x}_i)}{\max \phi(\mathbf{x}_i)} - \mathbf{y}_i}{\alpha - \mathbf{y}_i} \right\|^2 + \lambda \|\mathbf{1} - U\|^2$$

y_i are ground truth values and $\phi(x_i)$ are predicted values.

L_2 regularization term added to penalize the deviation of the prior mask U from its initial value.

Experimental Results

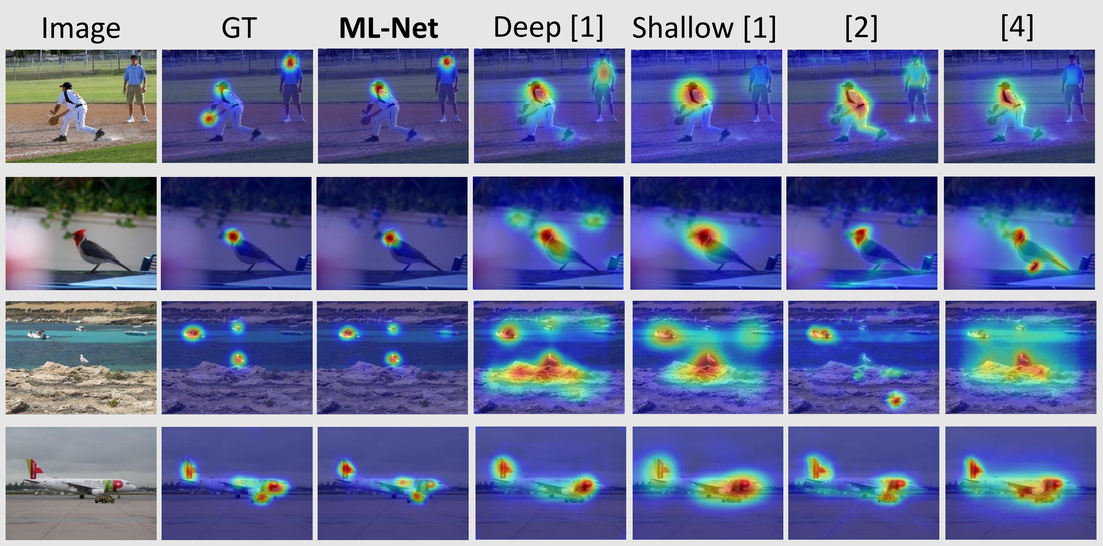
- We evaluate our model on the SALICON dataset and on the MIT Saliency Benchmark.

Results on SALICON Dataset

	CC	sAUC	AUC
ML-Net	**0.74**	**0.77**	**0.87**
Deep Convnet [1]	0.62	0.72	0.86
Shallow Convnet [1]	0.60	0.67	0.84
WHU IIP (LSUN Challenge 2015)	0.46	0.61	0.79
Rare 2012 Improved [2]	0.51	0.66	0.81
Xidian (LSUN Challenge 2015)	0.48	0.68	0.81
Baseline: BMS [3]	0.43	0.70	0.79
Baseline: GBVS [4]	0.42	0.63	0.79
Baseline: Itti [5]	0.20	0.61	0.67

Results on MIT300 Dataset

	Sim	CC	sAUC	AUC	NSS	EMD
Infinite humans	1.00	1.00	0.80	0.91	3.18	0.00
DeepFix [6]	0.67	0.78	0.71	0.87	2.26	2.04
SALICON [7]	0.60	0.74	0.74	0.87	2.12	2.62
ML-Net	**0.59**	**0.67**	**0.70**	**0.85**	**2.05**	**2.63**
Deep Convnet [1]	0.52	0.58	0.69	0.83	1.51	3.31
BMS [3]	0.51	0.55	0.65	0.83	1.41	3.35
Deep Gaze I [8]	0.46	0.51	0.76	0.87	1.29	4.00
Mr-CNN [9]	0.48	0.48	0.69	0.79	1.37	3.71
Shallow Convnet [1]	0.46	0.53	0.64	0.80	1.47	3.99
GBVS [4]	0.48	0.48	0.63	0.81	1.24	3.51
Rare 2012 Improved [2]	0.46	0.42	0.67	0.77	1.34	3.74



References & Code

[1] Pan, et al. "Shallow and Deep Convolutional Networks for Saliency Prediction." *CVPR*, 2016.
[2] Riche, et al. "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis." *SPIC*, 2013.
[3] Zhang, Jianming, and Stan Sclaroff. "Saliency detection: A boolean map approach." *ICCV*, 2013.
[4] Harel, Jonathan, Christof Koch, and Pietro Perona. "Graph-based visual saliency." *ANIPS*, 2006.
[5] Itti, et al. "A model of saliency-based visual attention for rapid scene analysis." *IEEE TPAMI*, 1998.
[6] Kruthiventi, et al. "DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations." *arXiv:1510.02927*, 2015.
[7] Huang, Xun, et al. "SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks." *ICCV*, 2015.
[8] Kümmerer, et al. "Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet." *arXiv:1411.1045*, 2014.
[9] Liu, Nian, et al. "Predicting eye fixations using convolutional neural networks." *CVPR*, 2015.

Code and project page
imagelab.ing.unimore.it
github.com/marcellacornia/mlnet