Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara: *SAM: Pushing the Limits of Saliency Prediction Models*; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 1890-1892

- Go beyond classical feed-forward networks, proposing SAM (Saliency Attentive Model). This incorporates neural attention mechanisms to iteratively refine predictions.
- Experiments confirm effectiveness and generalization capabilities of the model
- SAM
    - Saliency prediction models used for emulating where humans look in a scene useful for many computer vision applications (image captioning [M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Visual Saliency for Image Captioning in New Multimedia Services. In ICME Workshops, 2017] , auto cropping (M. Cornia, S. Pini, L. Baraldi, and R. Cucchiara. Automatic image cropping and selection using saliency: An application to historical manuscripts. In Dig)
    - Deep learning has greatly improved saliency prediction, as well as use of novel architectures and large datasets
    - Use of machine attention models rarely investigated in this task
    - SAM incorporates attentive mechanisms to iteratively define saliency predictions
    - SAM composed of 3 main components:
        - Dilated Convolutional Network – extracts feature maps from input image
        - Attentive Convolutional LSTM – recurrently enhances saliency features
        - Learned prior module – incorporates human-gaze centre bias in final predicitons
    - **Dilated Convolutional Network**
        - Deep saliency architectures usually built over pre-trained CNN that extracts feature maps from input images
        - Major drawback is that it drastically rescales image – worsening performance
        - Use of Dilated CNN limits rescaling effect – maps rescaled by factor of 8 instead of 32
        - Dilated convolutions and modifications of standard CNN architectures, produces saliency maps with an increased output size
    - **Attentive Convolutional LSTM (long short-term memory?)**
        - Recurrently processes saliency features at different locations
        - Extend traditional LSTM to work on special features by replacing dot product with convolutional operations – hidden states are feature stacks instead of vectors
        - Process features in 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; 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**
        - Output of LTSM combined with learned priors – used to model centre bias present in human-eye fixations
        - Network learns its own priors
        - Each prior is a 2d Gaussian function whose mean and covariance matrix are freely learnable
        - Priors are therefore inferred from data without relying on assumptions from biological studies
    - **Loss Function**
        - During training phase, network minimizes a combination of different cost functions taking into account different quality aspects that prediction should meet
        - Loss function is given by linear combination of 3 saliency evaluation metrics
            - Normalised scanpath saliency
            - Linear correlation coefficient
            - Kullback-Leibler divergence
        - These all commonly used to evaluate saliency prediction models
- Results
    - Tested on SALICON – 20,000 images with corresponding ground truths
    - Uses 2 versions of SALICON (recent one replaces velocity based fixation detection algorithm, resulting in more eye-like fixations) and compares results
    - SAM yields better results under all metrics except AUC where it loses out by 0.002 to DeepGazeII. SAM quantitively overcomes drawbacks of different existing proposals