Unsupervised Approach towards Visual Saliency Modelling

Homapriya Tarigopula, Beren Millidge, Richard Shillcock
School of Informatics, The University of Edinburgh
(s1793416@sms.ed.ac.uk , s1686853@sms.ed.ac.uk , rcs@inf.ed.ac.uk)

I. Introduction

Visual Salience refers to the most conspicuous regions of a visual scene. Visual salience modelling has gained enormous attention due to its wide range of applications in Robotics, Object segmentation and Object Tracking. Amongst the two variants of visual salience modelling, this work emphasises the bottom-up approach where the visual scenes or images are presented under free viewing conditions to track the salient regions. There have been several different approaches towards modelling visual salience starting from the earliest models based on feature extraction to the modern day deep learning. The recent deep learning models are supervised models which learn end-to-end from the input image to the saliency map. Though these models achieve reliable results, the major downside of these models is that they have very little or no relevance to the cognitive processes of the human brain. In this work, we propose an unsupervised learning model based on predictive processing [2] and information theory [1], to address the aforementioned drawback.

II. Methods

A. Concept

From the perspective of predictive processing, visual salience can be explained as follows: the brain is a generative process which continuously tries to predict the environment and adjusts its errors in prediction through observations involving ocular movements. The fixations of the eye are, in effect, guided along the most salient or conspicuous regions in the visual scene. This generative process of the brain is mimicked in our study through various architectures of Autoencoders. Since the reconstructed image by the autoencoders is similar to the generative process' prediction, it lacks salient information. Thus, the difference between the input image and the reconstructed image contributes towards saliency. Also based on information theory, the most salient information is observed to be least probable and hence the different brain regions fail in predicting these salient features accurately. Thus, the convolutional neural networks involved in the design of autoencoders also tend to lose salient information while reconstruction since these networks are inspired from mammalian visual cortical architectures.

B. Design

The choice of colour space was CIELAB similar to the Hering's opponent theory which explains the perception of colour in the human visual cortex. From our colour space, the luminosity (L) and chromaticity (a,b) channels are perceived as separate information. With the abstractions from these channels, our three models are described as follows:

- Observing Information on the whole: The information of L, a, b space is processed incrementally as a single joint information as depicted in Figure 1.
- Disjoint Processing of Information: The information from these discrete channels is considered independently while trying to learn each from other as shown in Figure 2.
- Jointly Processing the latent Information: The individual colour channels are encoded, and the latent information is jointly decoded as in Figure 3.

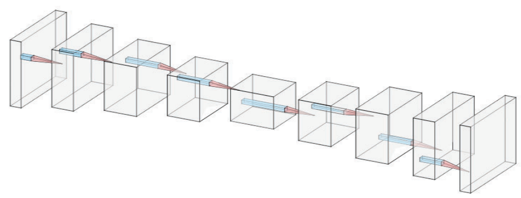

Figure 1: Observing information as whole

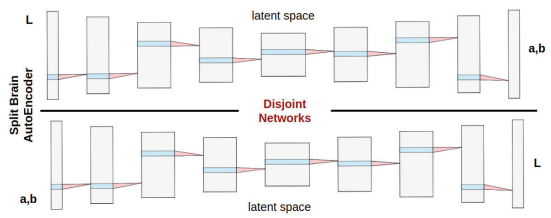

Figure 2 : Disjoint Processing of Information

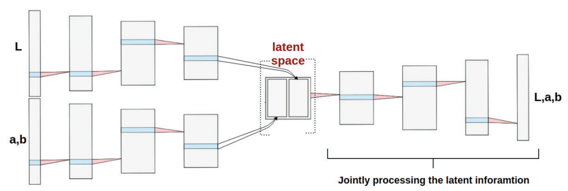

Figure 3: Jointly Processing the latent information

C. Implementation

A subset of test images from the IMAGENET dataset were used to pretrain the models. The data from the CAT2000 dataset [4] of the MIT Saliency Benchmark [3] was used to fine tune the learning model to reconstruct the input image. The different network architectures used as a part of testing out three different hypotheses were the traditional autoencoders (TA), the split brain autoencoders (SB) and an autoencoder network (JD) which disjointly encodes the data similar to the rods and cones in the eyes and jointly decodes them at a later stage from the latent space.

TABLE I. MODEL CONFIGURATIONS

TA	*SB*	*JD*
Input Channels : 3	Input Channels : [1 | 2]	Input Channels : [1 | 2]
Conv 3^1_{32}	Conv $3^1_{[16 \| 16]}$	Conv $3^1_{[16 \| 16]}$
MaxPool 2^1	MaxPool 2^1	MaxPool 2^1
Conv 3^1_{64}	Conv $3^1_{[32 \| 32]}$	Conv $3^1_{[32 \| 32]}$
MaxPool 2^1	MaxPool 2^1	MaxPool 2^1
Conv 3^1_{128}	Conv $3^1_{[64 \| 64]}$	Conv $3^1_{[64 \| 64]}$
MaxPool 2^1	MaxPool 2^1	MaxPool 2^1
Conv 3^1_{128}	Conv $3^1_{[64 \| 64]}$	Conv $3^1_{[64 \| 64]}$
Conv 3^1_{256}	Conv $3^1_{[128 \| 128]}$	Conv $3^1_{[128 \| 128]}$
T-Conv 3^2_{128}	T-Conv $3^2_{[64 \| 64]}$	T-Conv 3^2_{128}
T-Conv 3^2_{128}	T-Conv $3^2_{[64 \| 64]}$	T-Conv 3^2_{128}
T-Conv 3^2_{64}	T-Conv $3^2_{[32 \| 32]}$	T-Conv 3^2_{64}
T-Conv 3^2_{32}	T-Conv $3^2_{[16 \| 16]}$	T-Conv 3^2_{32}
T-Conv 3^2_{3}	T-Conv $3^2_{[2 \| 1]}$	T-Conv 3^2_{3}
Output Channels : 3	Output Channels: [2 | 1]	Output Channels: 3

In the **TABLE I** every convolution, transposed convolutional layer is defined with 3 parameters:$(\textit{size of filter})^{\textit{stride}}_{\textit{num of kernels}}$ For networks *SB* and *DA* $||$ is used to indicate disjoint-ness in the network architecture.

III. Results

We observe that the split-brain autoencoder could not reconstruct the image through cross channel prediction as the Luminance channel contains more information compared to the colour channels whereas the results observed qualitatively and quantitatively for the other two networks are presented below.

A. Qualitative Results

The difference image between the reconstructed image and the original image when averaged across the three channels and smoothed gives the saliency map as shown below:

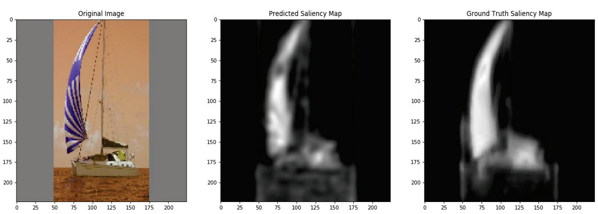

Figure 4: Comparison of outputs from TA with ground truth saliency map. From left to right, the original image, our predicted saliency map and the ground truth saliency map are shown.

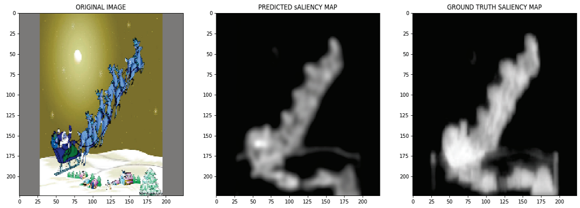

Figure 5: Comparison of outputs fromJD with ground truth saliency map. From left to right, the original image, our predicted saliency map and the ground truth saliency map are shown.

B. Quantitative Results

The models have been evaluated by the MIT Saliency Benchmark with the ground truth fixation locations of the observers to predict saliency. The evaluated results with respect to the different evaluation metrics are as presented below:

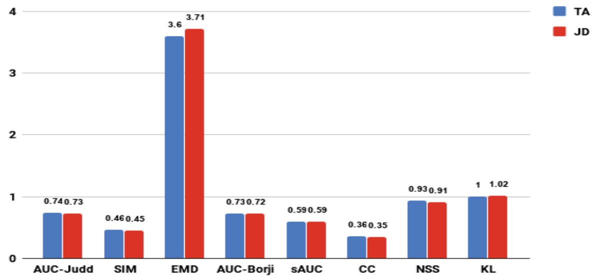

Figure 6: Quantitative Results

IV. Conclusions

A major contribution of these unsupervised models is that they developed an inherent centre-bias as depicted in Figure 7 unlike supervised models, which is a characteristic feature of saliency maps. The quantitative results produced by our models are better than feature based models but require further improvement to compare with other supervised deep learning based models.

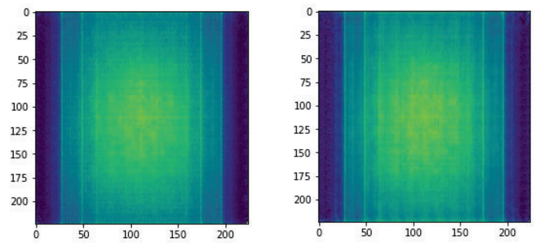

Figure 7: Centre-Bias left : TA, right : JD

References

[1] Attneave, F., 1954. Some informational aspects of visual perception. *Psychological review*, *61*(3), p.183.

[2] Friston, K., 2009. The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, *13*(7), pp.293-301.

[3] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A. and Torralba, A., 2014. Mit saliency benchmark (2015), p.13.

[4] Borji, A. and Itti, L., 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.