Image Saliency Prediction

G Pradyumn
2018CSB1088
Divyanshu Mathpal
2018CSB1086

Indian Institute of Technology, Ropar
Punjab, India

Abstract

Saliency is what is interesting or distinctive in a photo or scenario, enabling your eyes to quickly focus on the most important regions. Given an image, the brain selects part of the scene for further detailed processing, whereas the rest is discarded. It also prioritizes the data in such a way, that the selected parts are processed first.
This selection and ordering process is known as selective attention or visual saliency. The input in such tasks is an image and the output is a heat map image, displaying relative saliency of patches in the image.

1 Introduction

> *"The world is too much for us, it contains far too much information to be perceived at once"*
>
> *William Wordsworth*

Humans come across a large amount of visual data in their day to day life. If we start perceiving such a huge amount of data all the time, it would be difficult for us to perform even the easiest of tasks with accuracy. Visual attention enables humans to direct their resources into processing the most important segments of visual data. Intensive research has been carried out to reproduce our visual attention mechanism to object recognition, object tracking etc. Two different types of attention mechanisms have been identified [1]:-

1. Bottom up :- This attention mechanism seeks to help shift the focus to outliers, or the out-of-place patches of information.

2. Top Down :- This attention mechanism makes use of relevance of corresponding patches, and shifts focus accordingly.

Although attempts to replicate this saliency has been going on since the 1980s, there has been a radical increase in the accuracy and sophistication of saliency models from 2014 onwards. This can be attributed to large crowd-sourced saliency datasets, behavioural studies predicting gazes across images, and the rise of deep neural networks. They are accurate to such an extent that sometimes it is difficult to distinguish between human-annotated data and predicted data.

1.1 Scope

Here we will limit our discussion to the context of saliency prediction in images. Attributes related to saliency in videos[2], although closely linked to saliency in images, will not be discussed.

2 Datasets and evaluation metrics

2.1 Datasets

To annotate large scale datasets, researchers have resorted to crowd-sourcing techniques such as gaze tracking to webcams or mouse movements. Some of the most recent and influential image saliency datasets include MIT300, CAT2000, and SALICON datasets. An extensive list of datasets is given at the end of this subsection.

- **MIT300** :- This dataset consists of 300 highly varied images. It is annotated by 39 individuals ranging from ages 18-50. This is a particularly tough dataset for saliency prediction models. MIT Saliency Benchmark tests models against this dataset.

- **CAT2000[3]** :- This dataset was introduced in 2015. It consists of 20 different categories, and each image has a resolution of 1920 x 1080 pixels. This dataset consists of 2000 training images and 2000 testing images. MIT Saliency Benchmark tests models against this dataset. This dataset was annotated by 120 observers.

- **SALICON[4]** :- This dataset is by far the largest crowd-sourced dataset for saliency purposes, consisting of 15,000 training images, 10,000 training images, and 5,000 validation images. The images in this dataset come from the Microsoft COCO dataset. Currently, many datasets are first trained on SALICON dataset, then fine tuned on CAT2000 dataset. Then models are submitted to MIT benchmark or LSUN Saliency challenge for evaluation on their test sets.

2.2 Benchmarks

Benchmarks provide a sense of positive competition by publicly showing the rankings of submitted models according to various metrics. They played a key role in the advancement of saliency prediction models. Two of the most prominent benchmarks have been discussed here:-

- **MIT:[5]** The MIT benchmark is currently the gold standard in evaluating saliency prediction models. It tests the models against the MIT300 and CAT2000 datasets, and supports 8 different evaluation metrics.

- **LSUN Saliency Challenge:** This benchmark tests models against SALICON dataset. It supports over 7 different metrics and uses the same evaluation tools as MIT benchmark. Due to greater size of SALICON dataset it provides a better testing ground for saliency models.

3 Evaluation Metrics[6]

In general, they fall into two categories:-

- **Distribution-based** : Pearson's Correlation Coeffi- cient (CC), Kullback-Leibler divergence (KL), Earth Mover's Distance (EMD), and Similarity or histogram intersection (SIM)

- **Location-based** : Normalized Scanpath Saliency (NSS), Area under ROC Curve (AUC) and its variants including AUC-Judd, AUC-Borji, and Shuffled AUC (sAUC)

Both MIT and SALICON datasets use most of these metrics for their evaluation.

4 Models

4.1 Classical(non neural) models

The classic Itti[7] model used a bottom-up method to predict image salience. It can assign weights to different bottom-up features like color intensity, or orientation among others which changes the degree of priority of these features.
This model was not able to extract the higher level features or objects in images. These shortcomings paved the way for the use of Deep Neural Networks for predicting image saliency.

4.2 Deep neural network Models

CNN's have already achieved near-human accuracy in object detection tasks, which can be attributed to their ability to extract complex features from images. Saliency prediction models using such pre-trained CNN's for object detection tasks have thus proved to be much more accurate than traditional classical approaches, which were only able to extract low level, and very few high-level features. We have discussed state-of-the-art models here:-

- **EML-Net [8]** :- The major idea behind Expandable Multi Layer-Network is that the encoder and decoder parts of the architecture are trained separately, thus allowing for much deeper neural network architectures than VGG-16 and Resnet-50. It also provides a wide variety of possibilities of combining with other state-of-the-art architectures. In SALICON dataset, it outperformed all the other models in the NSS metric.

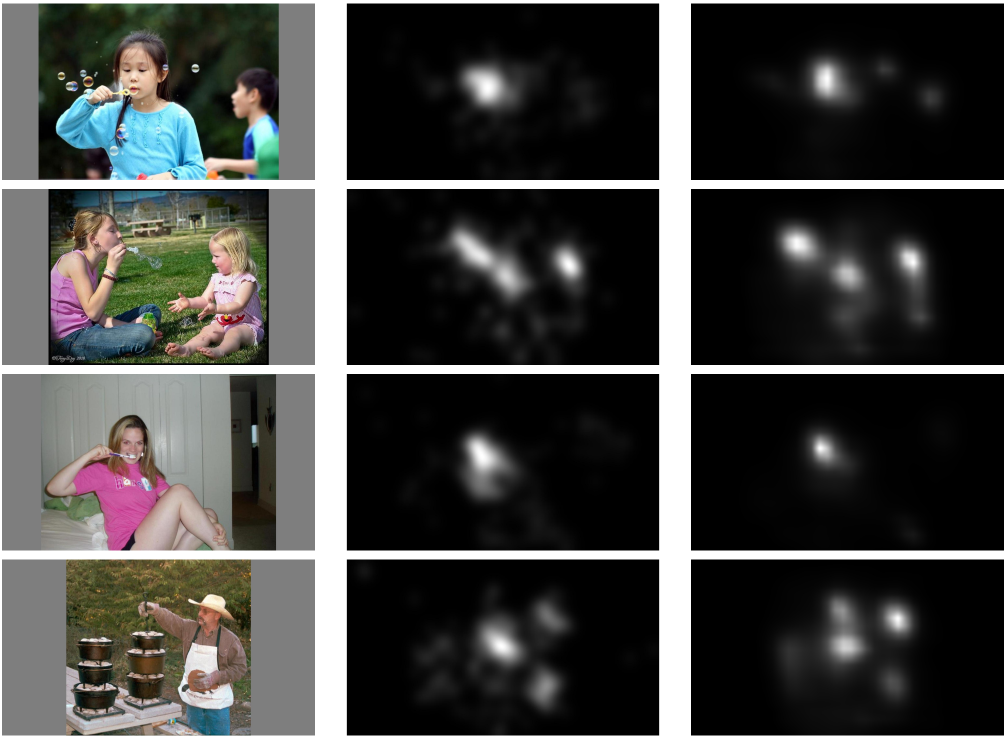
- **SAM-Net[9]** :- The most distinctive feature here was that instead of assuming a gaussian prior, it was learned from the dataset. It exploits an attentive convolutional LSTM architecture, where the LSTM module works on a stack of extracted features by the model. It also uses dilated convolution to increase the output resolution by just changing the strides in convolutional layers. It is one of the top models in SALICON and MIT benchmarks.

- **eDn [10]** :- The ensemble of Deep Networks (eDN) was proposed by Vig *et al* was the first model to use CNN's for image salience prediction. The model generates a large number of instances of a richly parameterized bio-inspired hierarchical models. Then it uses hyperparameter tuning to search for individual models and their combinations that are predictive of salience and then combines all these models into a single model using a linear classifier. At the time of its release in 2014, this model outperformed all the existing models on the MIT300 dataset.

- **Deepgaze II[11]** :- Deepgaze II was built upon Deepgaze I which used DNN features trained on object recognition (using AlexNet) and showed that it could outperform other models which trained deep features from scratch (like eDN). The outputs of these convolutional layers were used to create a model to predict image salience. The Deepgaze II model achieved the best score on the MIT300 AUC-Judd metric.

5 Preliminary Results

We have used a used Sam-ResNet model pre-trained on imagenet dataset for predictions here. The images were taken from CAT2000 dataset. Given below are the actual images, the ground-truth, and predictions by Sam-Net respectively for 4 images from this dataset.



Original Ground Truth Predicted

As we can see, the model predictions are not visually different from the ground truth annotations. We chose this model as our baseline network, as it introduced a lot of new ideas in saliency prediction, as well as has good accuracy on both SALICON and MIT benchmarks[9].

6 Our Approach

6.1 Background

We have built an AutoEncoder to identify the salient regions in an image. An AutoEncoder is a neural network consisting of an encoder and a decoder. The output of an ideal AutoEncoder is the same as its input. Thus an AutoEncoder tries to learn an approximation of the identity function

$$f(x) = x \tag{1}$$

. AutoEncoders have various uses ranging from data compression to de-noising images. There are 4 basic types of AutoEncoders[12]:

- Vanilla autoencoder
- Multilayer autoencoder
- Convolutional autoencoder
- Regularized autoencoder

The two most commonly used Regularized autoencoders are the sparse autoencoder and the de-noising autoencoder. We use the same principle as a de-noising autoencoder to predict the salient regions of an image by reducing background noise and focusing only on the relevant regions in the image.

6.2 Dataset

Initially we had planned to train a model on the salicon dataset, and then fine tune it on the CAT2000, but due GPU constraints in google colab we had to reduce the size of our dataset to 1,000 images in the training set and 500 images in the validation set. Initially, the images had a shape of (480,640,3) and the target saliency map was (480,640,1). We reduced their dimensions to (32,64,3) and (32,64,1) respectively.

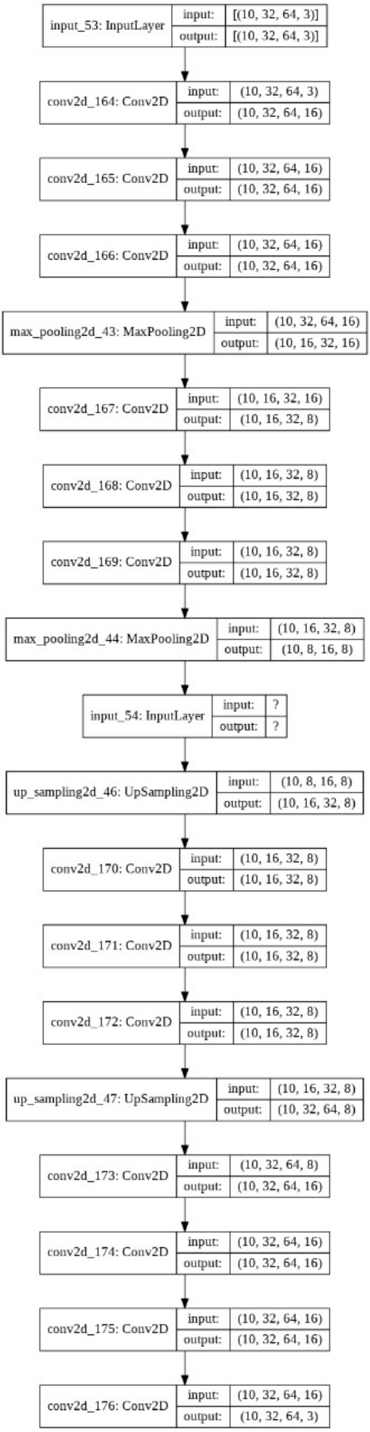
6.3 Preprocessing

The target saliency maps (of dimension (32,62,1)) were changed into RGB form. It was done as follows :-

The saliency map values were added to the Red component of images, which resulted in a reddish region over the salient regions. During earlier training it was observed that after around 40 epochs, the gradients started to explode, resulting in lower accuracy. Hence both the saliency maps and the input images were normalized.
The model was trained with a batch size of 10, and the gradient descent in case of loss function was rugged, as expected.

6.4 Architecture

Autoencoders usually require large number of epochs to learn useful features. The results we have shown were achieved after 300 epochs.



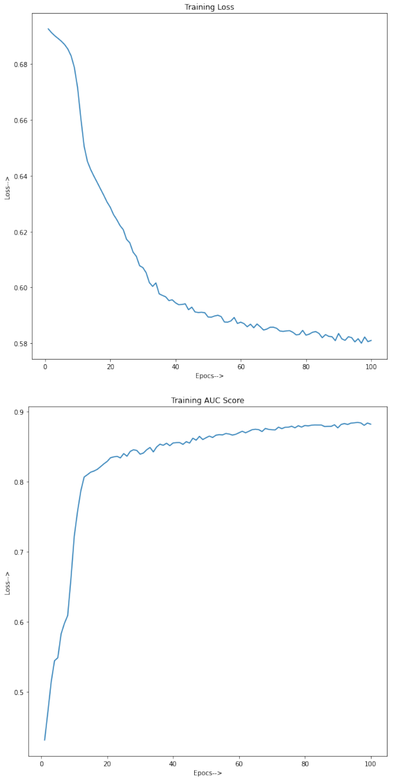
Custom CNN Architecture

The input will be the actual image and the output will be the saliency map for that image. If the number of output dimensions for the encoder is low, then the neural network will reduce the dimensions of the image by learning the salient regions of the image so that the decoder can reconstruct the original image with minimal loss in data. We allow the model to learn the importance of features based on the compressed representation of images.

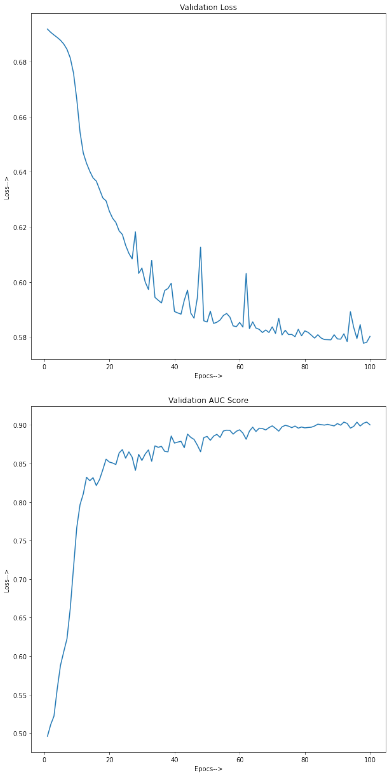
This is the basic principle that we have used to extract the salient regions of an image. The encoder and decoder we have used are of custom Convolution Neural Network Architecture. We had also tried VGG-16 based autoencoder, keeping the weights for the encoder part from imagenet dataset, and training the decoder on our dataset, but due very few images, and smaller dimensions on each of them, it was not able to learn anything, and the pixel values were approaching constant values.

6.5 Loss function and Metrics

We used binary cross-entropy as loss function with SGD(Stochastic Gradient Descent) as optimizer. AUC score was observed and used as a metric. Validation AUC score at the end of 300 epochs was around 0.92, after which the model had nearly achieved the local minima. Plots have been given for first 100 epochs.



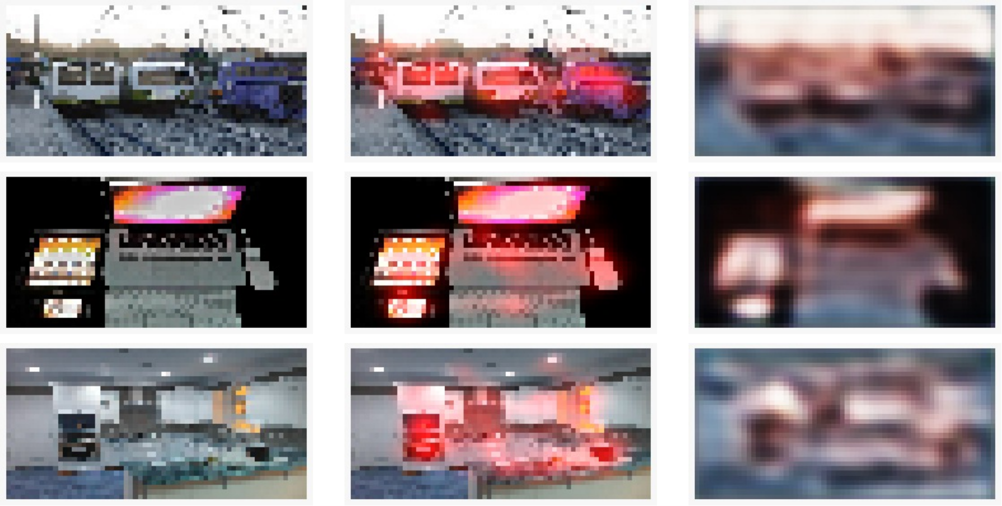
Training



Validation

7 Results

Given below are the actual images, the ground-truth, and predictions by our custom Autoencoder respectively for 3 images from the Salicon dataset.



Original Ground Truth Predicted

8 Further Work

Here, the size of dataset that we took was quite small, so with a larger dataset better results would be obtained. Futher images with larger dimensions would allow way for training much deeper networks, and thus acquire better scores. Such additions would require significantly greater hardware capabilities. Also, one can incorporate priors pre-trained on the dataset, for increasing performance.

References

[1] Charles E. Connor, Howard E. Egeth, and Steven Yantis. Visual attention: Bottom-up versus top-down. *Current Biology*, 14:850–852, 2004.

[2] Ali Borji. Saliency prediction in the deep learning era: Successes, limitations, and future challenges. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2018.

[3] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[4] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[5] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[8] Sen Jia. Eml-net:an expandable multi-layer network for saliency prediction. 05 2018.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[10] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014.

[11] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, 2017.

[12] Deep inside: Autoencoders. https://towardsdatascience.com/deep-inside-autoencoders-7e41f319999f. Accessed: 23-05-2020.