ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

VISUAL SALIENCY OBJECT DETECTION USING DEEP LEARNING TECHNIQUE

Rovina Mariam Jose
Department Of Computer science
Digital University
Kerala,India
Email:rovinajose95@gmail.com

Naveen Thomas Joseph
Department Of Computer science
Kristu Jyoti College
Kerala,India
Email:naveenntj@gmail.com

Abstract—Computer vision is a discipline of computer science that tries to mimic human vision systems. Visual attention is the subset of computer vision and it follows a systematic approach for the selection or concentration of specific information by neglecting the other perceivable information. This same aspect is deployed in salient object detection, through which a person's eye points to some visual scenes which in turn find some characteristics of the image. It drives to saliency, which measures how likely human eyes will fixate a certain part of the image and the part in which the eye focusing is the salient region of an image. The computer vision and image processing algorithms can be used to locate the most salient regions of an image and can create a saliency map. Saliency object detection aims to explore the salient part, by detecting and segmenting the object more clearly, it projects the outline of the salient object. This works aims to create a saliency object detection model based on fixation and segmentation. The saliency map that we proposed here can generate better results than existing models. This is achieved here by developing a model consists of three modules, generation of a mask of an image, segmentation of salient object .The mask and segments generated will be taken to find the mean of the feature vectors to create the saliency map. The performance of the system is done by the evaluation nss, auc, cc metrics and obtained 1.9,0.88,0.80 which is better than the existing salient object detection models.

Index Terms—Saliency map, R-CNN mask, VGG, Saliency Object Detection.

I. Introduction

The capacity to orient and maintain focus on a stimulus, such as a person, inanimate object, or task, is the aim of visual attention. The combination of bottom-up sensory input and top-down attentional cues forms an integrated saliency map according to the recent accumulating research.

The idea of visual attention is derived from one of the most fundamental capacities of the human visual system, the capacity to direct processing resources to important areas of sensory information. Human attention is guided by two systems, bottom-up and top-down mechanism. The bottom-up system directs attention to "salient" things, which appear to "pop out" of a picture due to some feature that distinguishes them from the rest of the picture such as color, orientation, or motion. Saliency detection is an important attentional skill that helps humans to direct their limited perceptual and cognitive resources onto the most relevant subset of available sensory input, enabling learning and survival. Top-down attention, on the other hand, directs attention to behaviorally important areas.

Fig. 1. Orginal Image

Fig. 2. Salient parts

II. Related Works

The [1] Saliency Using Natural statistics (SUN) model combines top-down and bottom-up information to predict eye movements during real world image search tasks. SUN implements target features as the top-down component. SUN outperformed a salience driven model in predicting human fixation positions during real-world image search. It uses natural

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

The [2] bottom-up visual saliency model, Graph-Based Visual Saliency (GBVS), is proposed. It consists of two steps, first forming activation maps on certain feature channels, and then normalizing them in a way which highlights complicity and admits combination with another map

The [3] The Spectral Residual Approach is intent to provide features, categories, or other forms of prior knowledge of the objects. By analyzing the log-spectrum of an input image extract the spectral residual of an image in spectral domain, and propose a fast method to construct the corresponding saliency map in spatial domain.

SalNet [4] approach uses a completely data-driven ap-proach, with a large amount of annotated data for saliency prediction. The model uses a CNN that consists of five layers with learned weights, three convolutional layers and two fully connected layers. Each of these three convolutional layers is followed by a rectified linear unit non-linearity (ReLU) and a max pooling layers. The deconvolution layer follows the final convolution to produce a saliency map that reflects the input width and height.

SalGAN [5] consists of two networks, one predicts saliency maps from a raw pixels of an input image where as the other one takes the output of the first one to discriminate whether a saliency map is a predicted one or ground truth. It contains a generator and a discriminator networks. The filter weights in SalGAN have been trained over a some loss resulting from combination of a content and adversarial loss. The content loss is calculated in a perpixel basis, where each value of the predicted saliency map is compared with its corresponding peer from the ground truth map.

DeepFix [6] a fully convolutional neural network for ac-curate saliency prediction. The input image is followed by 5 convolutional blocks. It is similar to the VGG-16net. These are the followed by an inception block and consists of a set of convolution layers with different kernel sizes operating in parallel. The weights of VGG-16 network have been learnt by training on 1.3 million images of the ImageNet database. These are followed by a Location Biased Convolutional (LBC) layers.

EML-NET [7] is a Expandable multi-layer network Multiple powerful deep CNN models to better extract visual features for saliency prediction. The DenseNet-161 mode list pre-trained on the PLACE365 dataset and the NasNet-Large model is pretrained on ImageNet. The input image applied to the CNN model to extract the feature map. The encoder and decoder are separately trained.

III. METHODOLOGY

Visual attention is the process that helps the human visual system to select the most relevant information from a visual scene. Deep neural networks have provided great success in image saliency prediction. The capacity to locate items or regions reflecting a distinct scene is referred to as saliency detection, simulating scenarios in which observers are see-ing their surroundings without a defined aim. A variety of computational models are being built, and the technique is being improved. Initially, models are created with the goal of forecasting human visual attention. The visual process of a person in an image or in a scene by a free moving condition is used to measure performance.

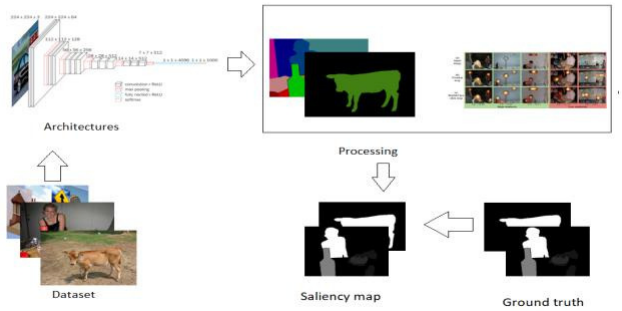

Fig. 3. Basic Structure

The figure 3 shows the structure of the saliency object detection process, there are several datasets available for this process. Each image is obtained from an eye tracker device, with the most often used datasets being MIT300, SALICON, PASCAL-S, MIT1003, and so on. Each image may have a fixation point, which is the place on which our eyes fixate for a few seconds over the image, as well as the associated ground truth, which is utilized to determine the calculation's loss. Each dataset contains images for training, testing, and validation. Following the validation of the datasets, the next stage is to construct a convolutional neural network model through which a featuremap is generated. There are serval types of CNN networks like AlexNet, VGG, ResNet. The pooling layers, which resize the images, are one of the key drawbacks of neural network architectures. Because the focus might be anywhere on the image, shrinking the image size may results in inaccuracies. As a result, the pooling layers after the blocks will be removed. After processing the images through the network, the very next phase is to discover the techniques 18 for generating the saliency map of the images, such as minimal barrier techniques, fine grained approaches, and mean of the feature vectors, segmentation and fixation etc. Finally evaluate the model with evaluation matrixes

Initially load a set of images as the dataset, and subdivides them into training, testing and validation sets. In addition, the dataset also carries videos and convert them into frames. Using any one of the architectures of CNN we extract the features of the images and created a feature map of the image. Since its aim is to detect the salient object, we use an faster R-CNN approach and segments and grab cut the masked images. Calculate the fixation point using the python library scades . The saliency map is the combination of fixation point and the segmentation and also taking the threshold of the images. Calculating the loss , evaluation metrics, F-measure with the help of ground truth images.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

A. CNN

Convolutional neural network is composed of multiple building blocks, such as convolution layers, pooling layers, and fully connected layers, and is designed to automatically and adaptively learn spatial hierarchies of features through a backpropagation algorithm. They are used mainly for image processing, classification, segmentation and also for other auto correlated data. There are different types of network in CNN like RESNET, VGG16, IMGNET,ALEXNET etc. are a special kind of multilayer neural networks, designed to recognize visual patterns directly from pixel images.We build our architecture on the popular 16 layers model from VGG, which known for its elegance and simplicity, and at the same time yields nearly state of the art results in image classification and good generalization properties. We build our architecture

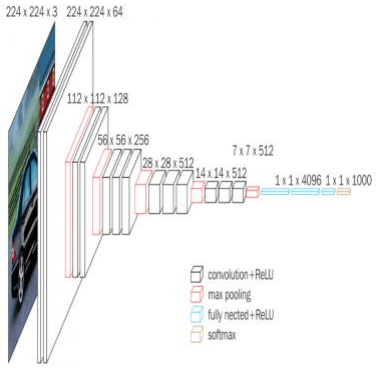

Fig. 4. Vgg Architecture

on the popular 16 layers model from VGG, which known for its elegance and simplicity, and at the same time yields nearly state of the art results in image classification and good generalization properties.

B. SEGMENTATION

Image segmentation is the technique of dividing a digital image into several parts in digital image processing and computer vision. The purpose of segmentation is to simplify and/or transform the representation of a picture into something more relevant and easier to examine. We focus largely on object identification and segmentation for object recognition using computer vision algorithms.

Pass the image through selective search and generate region proposal. R-CNN detection system consists of three modules. The first generates category-independent region proposals. These proposals identify the set of candidate detections present in an image. The second module is a deep convolutional neural network that extracts a feature vector from each region. Each image is segmented 25 and partial form masks are created. These masks contain information on the entire objects in the supplied image. The end outcome of this approach is accomplished as masks of detected objects that are compared to the ground truth with the use of a pixel-wise log loss function. The produced shape encodes the spatial arrangement of the input image full-connected layers translate class labels and bounding boxes into a feature vector from which the spatial structure of produced shapes may be derived

C. SALIENCY MAPS

Saliency Map is a picture in which the brightness of a pixel shows how salient the pixel is. The brightness of a pixel is directly proportional to its saliency. It is mostly a greyscale picture. Saliency maps are also called as a heat map where hotness refers to those regions of the image which have a big impact on predicting the class which the object belongs to.

The retrieved characteristics, such as color, orientation, and intensity, are transformed to feature maps in the form of vectors using the VGG-16. A proper representation for salient object segmentation, as well as computational concepts of feature saliency, such as area contrast, are two components of salient object algorithms. Here saliency is the combination fixation point and the segmentation of the image. and finding the mean vector and taking the binary threshold of the image we get the saliency map.

IV. RESULT

A. DATASET

In this work we are using pascal S dataset contains 850 images natural images from both indoor and outdoor scenarios. Collected on 8 subjects, 3s viewing time, Eyelink II eye tracker. The performance of most algorithms suggest that PASCAL-S is less biased than most of the saliency datasets

B. RESULT AND DISCUSSION

Let us Consider a CNN with a one convolutional layer which is followed by a fully connected layer trained to predict fixations. This model generalizes the classic model and that in turn learn to combine feature maps. The learned features in the CNN can be combined linearly by the fully connected layer. In order to handle the scale dependency of saliency computation, the classic models often recruit multiple image resolutions. But the deep saliency models concatenate maps from several convolutional layers, or feed input from different encoder layers to the decoder to get the det preserve fine details. The most evident difference of classic models compares to deep architectures is the lack of ability to extract higher level features, objects, or parts of objects. The deep structure of CNNs allows capturing complex features that attract gaze automatically. This is the main reason behind the big performance gap between the two types of models.

C. Evaluation Metrics

The main purpose of evaluation metrics is to compare the ground truth map with saliency map generated from the model.There are many different types of evaluation metrics available to test a model.y.Saliency prediction results are usually evaluated with a large variety of metric Commonly used Evaluation metrics, including Earth Movers Distance (EMD), Normalized Scan path Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd,

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

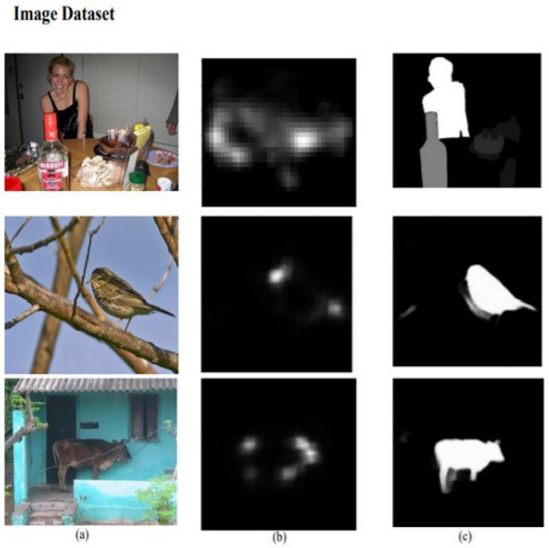

Fig. 5. (a) original image (b) previous model saliency map (c) new model saliency map

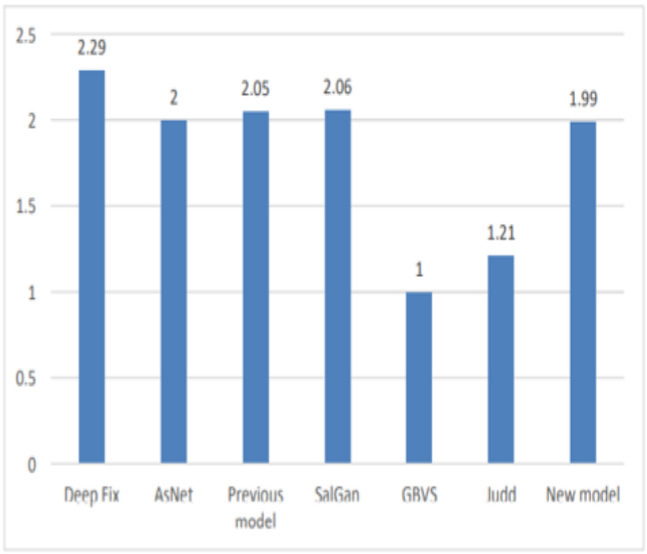

Fig. 6. Normalized Scanpath Saliency graph over different model

AUC-Borji, and shuffled AUC. In our experiments we use AUC Judd, AUC Borji and shuffled AUC. The AUC Judd and the AUC Borji choose non-fixation points with a uniform distribution, otherwise shuffled AUC uses human fixations of other images in the dataset as non-fixation distribution.

1) Earth Movers Distance: EMD is a measure of the distance between the two 2D maps. It is the minimal cost of transforming the probability distribution of the estimated saliency map to that of the ground truth map .

2) Normalized Scanpath Saliency (NSS): This metric is calculated by taking the mean of scores assigned by the unit normalized saliency map . When NSS value is 1 the saliency map shows significantly higher saliency values at human fixated locations compared to other locations. When NSS 0 indicates that the model performs no better than picking a random position, and hence is at chance in predicting human gaze.

3) Area Under Curve (AUC): In AUC, there are two image locations that are used actual human fixations as the positive set (fixation distribution) and some points randomly sampled from the image as the negative set. Depending on the choice of the non-fixation distribution the AUC are classified into 2 types. AUC with uniform distribution of non-fixation points (AUC-Judd and AUC-Borji) and the shuffled-AUC. The saliency map S is then treated as a binary classifier to separate the positive samples upon the negatives. Thresholding over the saliency map and plotting true positive rate on the false positive rate an ROC curve is achieved and its underneath area is calculated.

4) Linear Correlation Coefficient (CC): It measures how correlated or dependent two variables are and CC can be used to interpret saliency and fixation maps, G and S as random variables to measure the linear relationship between

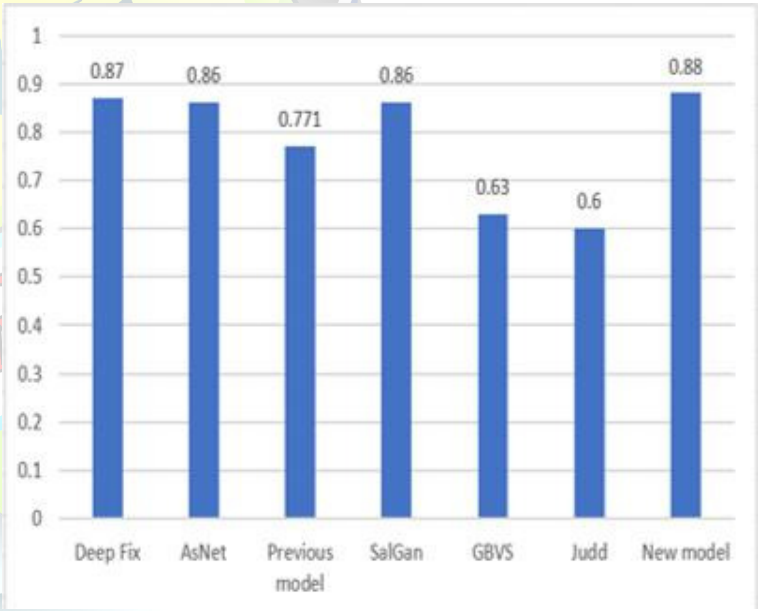

Fig. 7. Area Under Curve graph over different models

them,cov(S, G) finds the covariance of S and G and ranges between -1 and +1, and a score close to -1 or +1 indicates a perfect alignment between the two maps. The most evident difference of classic models compares to deep architectures is the lack of ability to extract higher level features, objects, or parts of objects. The deep structure of CNNs allows capturing complex features that attract gaze automatically. This is the main reason behind the big performance gap between the two types of models.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

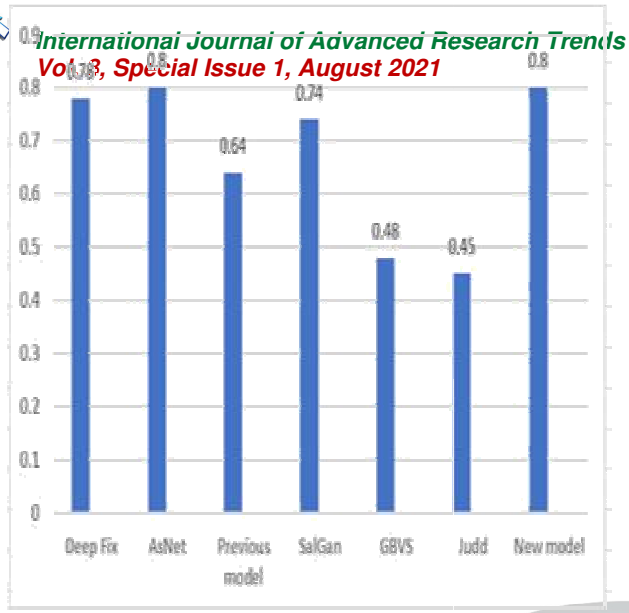

Fig. 8. Linear Correlation Coefficient graph over different models

V. Conclusion

Vision in general and images in particular have played an important role in human life. Visual saliency has been an increasingly active research area over the last few years. Several models are developed to find the visual gaze in an image. These are commonly represented by saliency map. This work presented a novel deep learning architecture for-saliency prediction. Our model uses segmentation and fixation approaches to generate the saliency map. The map generated point out the salient object more clearly than other model. And calculate the AUC, CC and NSS and get a value about 0.88,0.80 and 1.9.

References

[1] SUN: A Bayesian framework for saliency using natural statistics Lingyun Zhang; Matthew H. Tong; Tim K. Marks; Honghao Shan; Garrison W. Cottrell
[2] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” vol. 19, 01 2006, pp. 545– 552
[3] Hou, Xiaodi Zhang, Liqing. (2007). Saliency Detection: A Spectral Residual Approach. IEEE Conference in Computer Vision and Pattern Recognition. 2007. 10.1109/CVPR.2007.383267.
[4] Han, Le Li, Xuelong Dong, Yongsheng. (2019). Convolutional Edge Constraint-Based U-Net for Salient Object Detection. IEEE Access. 10.1109/ACCESS.2019.2910572.
[5] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O’Connor, Jordi Torres, Elisa Sayrol and Xavier Giro-i-Nieto. ”SalGAN: Visual Saliency Prediction with Generative Adversarial Networks.” arXiv. 2017.
[6] Rahul Gupta, Soham Pal, Aditya Kanade, Shirish Shevade DeepFix: Fixing Common C Language Errors by Deep Learning,
[7] Sen Jia, Neil D.B. Bruce, EML-NET: An Expandable Multi-Layer NETwork for saliency prediction, Image and Vision Computing, Volume 95, 2020, 103887,
[8] T. Judd, F. Durand, and A. Torralba, “A benchmark of computational models of saliency to predict human fixations,” in MIT Technical Report, 2012
[10] N. Bruce and J. Tsotsos, “Attention based on information maximization,” Journal of Vision - J VISION, vol. 7, pp. 950–950, 06 2010
[11] V. Leboran,´ A. Garc´ıa-D´ıaz, X. R. Fdez-Vidal, and X. M. Pardo, “Dynamic whitening saliency,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 5, pp. 893–907, 2017.
[12] J. Zhang and S. Sclaroff, “saliency detection: a Boolean map approach,” in Proc. of the IEEE International Conference on Computer Vision (ICCV), 2013.
[13] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, “Learning to detect salient objects with image-level supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017.
[14] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling:Salient Object Detection in the Deep Learning Era: An In-Depth Survey. CoRR abs/1904.09146 (2019)
[15] W. Wang, J. Shen, X. Dong and A. Borji, ”Salient Object Detec-tion Driven by Fixation Prediction,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711-1720, doi: 10.1109/CVPR.2018.00184
[16] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O’Connor, “Shallow and deep convolutional networks for saliency prediction,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016