Learning Salient Objects in a Scene using Superpixel-augmented Convolutional Neural Network

Shashank Tripathi, Yash Patel

***Abstract*—In this project, we are trying to determine salient objects in a scene. Salient objects are loosely defined as objects that pop-out in a scene in the opinion of a human observer – an object that attracts attention of the human brain and visual system.**

Many previous studies have attempted to determine saliency in an image. This project specifically attempts to use Convolutional Neural Network to solve the binary labelling problem, where salient objects are labelled 1 and background is labelled 0. An obvious problem with using CNN directly is that CNNs use a highly low-resolution image as input (227x227 pixels for the popular Alexnet). Contrast based saliency should be detected from a larger context and such low resolutions put a cost on the context in the image. A solution is to cluster the image into Superpixels, which retain long-range context, while maintaining low-resolution input. Superpixel suffer from loss of spatial information, so we try to reinject spatial information by the incorporating the following observations:

- **It has been shown in previous studies that contrast is a major factor to determine visual attention among humans**
- **Another factor that determines visual attention is if the object is in the image foreground. Objects with contrast can occur scattered across the background, but foreground objects usually tend to be locally connected**

***Index Terms*—Saliency Detection, Convolutional Neural Networks, SLIC superpixels**

I. Introduction

Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. On the Origin of Species, Charles Darwin.

Just as Charles Darwin attributed the success of all life on earth to a ubiquitous power called Evolution, we strongly believe that humans are only a biproduct of years of trial-and-error. The human visual system has developed over the years to optimally trade-off energy consumption and utility. It has been shown that the human visual system processes images in a hierarchical fashion [Marr and Poggio paper, 1982], starting from low-level context independent features like edges, contrast variation, color separation, etc to high-level features such as shape and structure. This observation has been instrumental in the development of the Convolutional Neural Networks which learn hierarchical weights in a similar fashion. This paper focuses on the use of low-level features to model human attention in a natural image.

Nassier et al (1964) proposed a model of the human visual system consisting of pre-attentive and attentive stages. The pre-attentive stage focuses human attention on pop-out features in the image. These pop-out features [Julesz, 1995] are regions in the image that present some form of spatial discontinuity, be it contrast separation or color variance. These features define Visual Saliency. Saliency detection is therefore an attempt to mimic the human visual systems pre-attentive stage. In the context of this project, we try to use the power of convolutional neural networks to identify salient regions in an image.

Many past approaches at saliency detection try to capture contrast cues to determine salient regions [Cheng et al 2011, He and Lau 2014]. However, just as computer vision approaches using hand-crafted features have been shown to be unsuitable for a general setting, these approaches too fail to generalise to all images. Many researchers have therefore tried to adapt state-of-the-art learning techniques to detect salient regions but have primarily focused on integrating saliency maps obtained from hand-crafted features. jiang et al, 2013]. This project takes inspiration from the results presented by the work of He et al, where they have tried to extract saliency labels on superpixels using a shallow Convolutional Neural Network. Using superpixels allow to reduce input dimension while retaining large image context.

Our approach:

II. Dataset

To train the Convolutional Neural Network, we have used the Extended Complex Scene Saliency dataset (ECSSD)[Heirarchical Saliency Detection, Yan et al] which provides 1000 natural images, along with ground truth labels. A few examples are shown in 1.

III. Methodology

An idea could be to directly feed the image pixels as input to a CNN architecture and generate saliency labels for each of the pixels. The issue with this approach is that most CNN architecture heavily downsample the input image (227x277 pixels for Alexnet, 224*244 pixels for VggNet and GoogleNet). Such a small image patch is insufficient to detect saliency as saliency detection typically needs a larger image context. Low resolution puts a cost on the larger context in

Fig. 1. Few sample image and corresponging ground truth labels from the ECSSD

the image. The convolution operation also occurs in a 3x3 neighbourhood so as to capture local dependencies. While this might be useful for applications like image classification/ segmentation, which primarily rely on local information, saliency information is embedded globally. Local convolutions operation therefore fail to capture the global image context if directly applied on raw pixel values.

To mitigate the above issue, we propose using superpixel segmentation on the image. Humans dont look at an image as a discretised matrix of pixels. To bridge the gap between how the human visual system perceives an image, many studies have proposed grouping pixels into perceptually meaningful regions that capture image redundancy. In general, superpixel regions have the following properties:

1) Superpixels fit well to the object boundaries in the image
2) Reduce pixelwise redundancies by grouping local pixels together based on color and intensity similarities
3) Superpixels should be efficient to compute and should offer an efficiency advantage over and above their computation cost when used in a subsequent task

Achanta et al have forwarded a powerful algorithm for superpixel segmentation called Simple Linear Iterative Clustering (SLIC). SLIC superpixels are based on the k-means clustering algorithm at its core. A simple modification that SLIC uses over the nave k-means clustering is that it restricts the search space to a neighbourhood of the current mean. The neighbourhood is proportional to the superpixel size. This reduces the computational complexity to linear time (O(N)) where N are the number pixels are independent of the number of superpixels. Instead of taking Euclidean distance to propose cluster assignment, a color and spatial proximity based distance measure is used instead. The algorithm for SLIC superpixels can be summarised as below [Achanta et al]:

1) Convert images from RGB colorspace to LAB colorspace

   CIE LAB colorspace bears a closer resemblance to human perception. It treats black and white as their own channel, i.e. it separates contrast from color, thereby having the advantage of a wider gamut. Contrast and color separation is important especially with regards to saliency detection as salient objects are localized based on both color and contrast separation in the context of the full image.
2) Write the rest of the algorithm from the slic superpixel paper. Describe the distance measure

?? shows a few examples of superpixel segmentation on text images. It is apparent that the superpixels accurately group pixels based on color similarities. Moreover, superpixels adhere well to image boundaries.

An inherent problem with superpixel segmentation is that each time a different set of superpixels are generated stochastically depending upon the initial initialization of the cluster means. In the process, Superpixels lose spatial information. To recover spatial information, a color uniqueness matrix can be defined on the superpixels [He et al]. Let a give image I and the segmented regions $R = [r_1, \ldots, r_x, \ldots, r_N]^T$, a $N \times N \times 3$ color uniqueness matrix Q can be defined.

$$Q = \begin{bmatrix} q^c_{11} & \cdots & q^c_{1j} & \cdots & q^c_{1M} \\ \vdots & \ddots & & \ddots & \\ q^c_{x1} & \cdots & q^c_{xj} & \cdots & q^c_{xM} \\ \vdots & \ddots & & \ddots & \\ q^c_{N1} & \cdots & q^c_{Nj} & \cdots & q^c_{NM} \end{bmatrix}$$

where each element q^c_{xj} represents the weighted difference in the mean color vector (across channels) of superpixel region x with every other superpixel region j. So $M = N$ in our case.

$$q^c_{xj} = t(r_j).|C(r_x) - C(r_j)|.w(P(r_x), P(rj))$$

where $t(r_j)$ counts the total number of pixels in region r_j. This term weights superpixels with more number of pixels to have higher contribution to the contrast than those with fewer pixels. $C(r_x)$ is the mean color vector of region r_x and $|C(r_x) - C(r_j)|$ is the 3D vector storing absolute differences of each color channel. $P(r_x)$ is the mean position of (r_x). The term $w(P(r_x), P(rj))$ is a Gaussian weight to attach higher contribution to pixels spatially closer to each other and $w(P(r_x), P(rj)) = exp(-\frac{1}{2\sigma_s^2}||P(r_x) - P(r_j)||^2$. Each row in Q is then sorted by the spatial distance to region r_x to maintain local correlation such that the convolution operation in the CNN makes sense. Sorting groups neighbouring superpixels together. In summary, each row of the Q matrix describes the color differences between each region r_X with all other M-1 superpixels in the image. The Gaussian distance weights include spatial information into the Q matrix which would have been lost otherwise.

The color uniqueness matrix does a good job in capturing salient objects separated by color variations in the image. Q matrix essentially represents regions that show significant color rarity in their neighbourhood. However, considering just color rarity to determine saliency might be an oversimplification. Consider images in 2 and their ground-truth salient object segmentations. It is apparent that even though certain objects display color rarity in the image eg. flowers, colored

balls in the background, fishes etc, they are not salient. We need to ignore superpixels that display higher local contrast separation but form part of the background. Primarily, we need to a metric to differentiate foreground objects from background objects. [Lie et al, 2011] show that foreground objects are compact i.e. locally connected, whereas background objects are distributed in the image. To capture this difference, we define a new matrix Q' which is complementary to the original Q matrix.

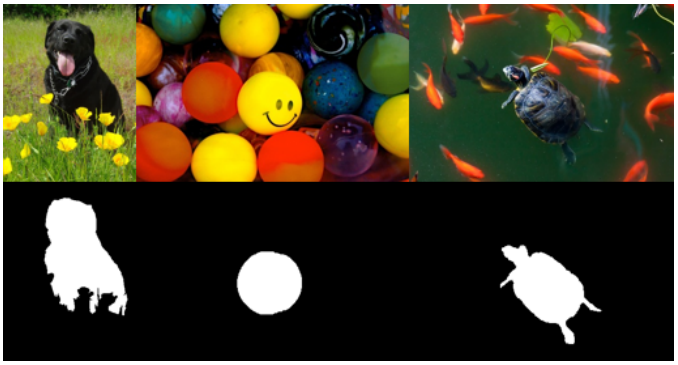

Fig. 2. Example of scattered non-salient background

$$Q' = \begin{bmatrix} q^d_{11} & \cdots & q^d_{1j} & \cdots & q^d_{1M} \\ \vdots & \ddots & & \ddots & \\ q^d_{x1} & \cdots & q^d_{xj} & \cdots & q^d_{xM} \\ \vdots & \ddots & & \ddots & \\ q^d_{N1} & \cdots & q^d_{Nj} & \cdots & q^d_{NM} \end{bmatrix}$$

where each element q^d_{xj}; represents the weighted difference in the mean position vector (across channels) of superpixel region x with every other superpixel region j. Each q^d_{xj} can be defined as:

$$q^d_{xj} = t(r_j).|P(r_x) - P(r_j)|.w(C(r_x), C(rj))$$

where $t(r_j)$ counts the total number of pixels in region r_j. This term weights superpixels with more number of pixels to have higher contribution to the contrast than those with fewer pixels. $P(r_x)$ is the mean position vector of region r_x and $|P(r_x) - P(r_j)|$ is the 3D vector storing absolute differences of each color channel. $C(r_x)$ is the mean color of (r_x) across the 3 channels. The term $w(C(r_x), C(rj))$ is a Gaussian weight to attach higher contribution to pixels spatially closer to each other in the color space and $w(C(r_x), C(rj)) = exp(-\frac{1}{2\sigma_s^2}||C(r_x) - C(r_j)||^2$. Each row in Q' is then sorted by the spatial distance to region r_x to maintain local correlation such that the convolution operation in the CNN makes sense. This is similar to what we did while forming the Q matrix.

We look at saliency detection as a binary labeling problem where superpixels are classified as either salient or not. The relationship between the binary labellings and the Q matrix can be efficiently learned by a shallow Convolutional Neural Network. Salient objects display a perceptible difference in color compared to the surrounding and are locally connected. Our hypothesis is that the Q matrix accurately encompasses these properties of salient object from which a convolutional neural network can accurately learn internal representation to classify previously unseen images.

IV. Network Architecture

In this work, we use a shallow 3 stage Convolutional Neural Network with each stage having a 1D convolutional layer, a max-pooling operator, and a RELU activation. The first two stages include [] filters with kernel size 50 each. The architecure is summarized in **??**.

V. Experiments