Combining class conditioned representations in CNN for Salient Object Segmentation

Anonymous ACPR submission

Paper ID ****

Abstract

Saliency detection is a key component of frameworks which attempt to solve challenging problems in Computer Vision such as Video Compression, object tagging, etc. In recent times, Convolutional neural network (CNN) has arisen as a promising framework in accurate modeling of visual saliency. This paper involves development of a CNN Framework followed by Dense Conditional Random Field (CRF) for Saliency Detection. It attempts to understand effect of inception like module for fusing class conditioned representations along with provision of alternate pathway for back-propagation. It incorporates this artificial bias in CNN , to diversify and then cohesively combine multilevel representations branching out of inception based scale normalized pooling responses.

1. Introduction

Among various striking features present in Human Vision,it's ability to discriminate and lay emphesis[19] on some region over other,distinctly sets it apart.Study[6] on lateral intra-pariatal area(LIP) in Macaque monkeys, responsible for abruptly appearing retinal simuli show that it's underlying visual representation is sparse, largely consisting of most salient or behaviorally relevant objects.Modelling this "focus of attaintion" is of great significance,not only for getting insight into human vision,but is also of a great interest for foveated Video Compression[9]. Work by Itti[10] and many others[15][8] exploit these biological cues for modeling visual attaintion.These models[1] take into account the ability of primates to combine interaction of preconcieved memory present in higher cortical areas with raw image information present in visual cortex. Convolutional Neural Network stands out as a natural choice,when it comes to Study of human fixations. It inherently takes in account, neuronal division of vision task , indicated by differences in spatiotemporal registration. Salient Object Segmentation deals with separating salient object from background. This task can be broken up into two parts namely saliency map generation and feature engineering. For instance, work such as SALICON[12] aims to generate saliency map using mouse contingent tracking. It selects salient parts out of image using part based Microsoft COCO dataset. CAT2000 dataset [2], generates eye contingent tracking for generating fixation maps. The second paradigm of research deals with feature engineering, given a saliency map. Simonyan Et Al.[17] focusses on engineering the model itself in order to have better classification and localization in Imagenet[16] challenge. It is important to observe that these series of approaches[17][12][2] aim to combine multiple representation, be it parts, regions (RCNN- [5]),segments,etc.Neuro-Cognitive interactions
Recent success of Convolutional Neural Networks for various computer vision tasks[3][17] indicates the drift of focus, from paradigm of devising stratagies to innovate new features and techniques of combining them, to a paradigm of engineering the very structure of convolutional neural network. Concatination of differently sized filter responses in CNN, was an idea, introduced in GoogleNet[18] by Szegedy Et Al. This block of differently sized filters along with concatination layer is called Inception module in GoogleNet[18] which aims to bring scale invarience at different depths of CNN. We include this module at various depths, however unlike googlenet[18], it is not a part of main branch of CNN, but is rather applied to the response of each pool layer in main branch.

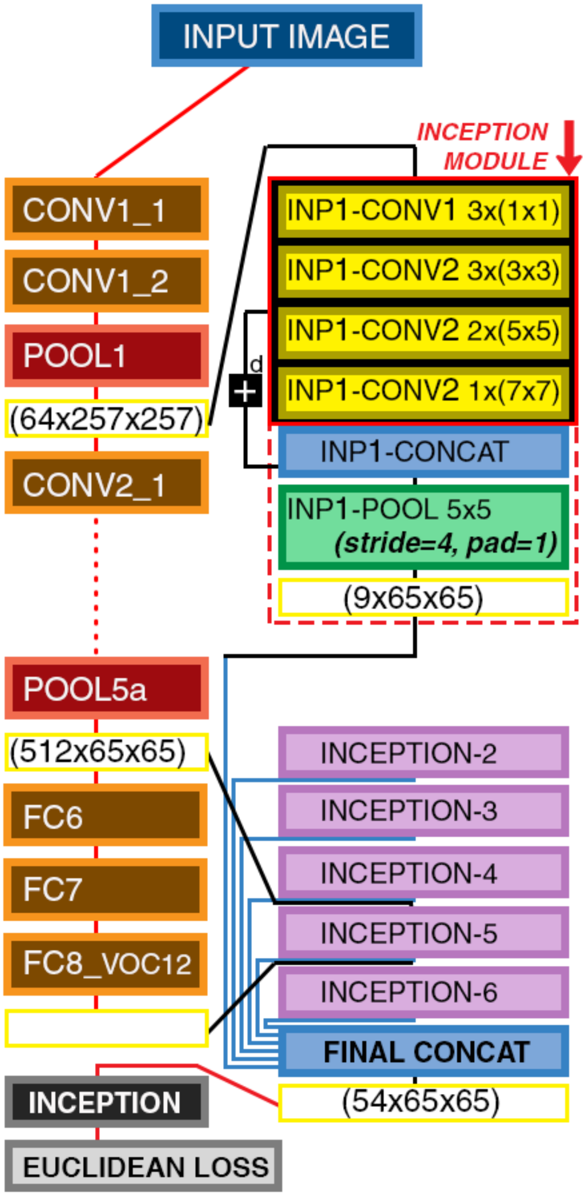
Next step consists of refining this saliency map to obtain a final binary segmentation map. Unlike Harel Et Al,[7], which uses markov chain to refine saliency based on context information, dense CRF[13] inherently takes care of context through fully-connected nature of pairwise potential.

Our approach is a combination of supervised CNN Modelling and unsupervised DenseCRF Modelling. Convolutional Neural Network training is a learning algorithm, which recursively convolves(or operates) intermediate responses to generate final heatmap. The weights of these filters are learned via stochaistic gradient descent. We use C++ based Caffe[11] and one of it's python based wrapper Expresso[4] for training these weights. The layers of CNN, consists of series of convolutions and pooling. Output of these auxialary branches spurring from output of pool layers is refined by inception modules are thereby concatenated together, forming a feature pyramid. This is again combined via inception like module in order to obtain scale invariant multilevel representation. Further, Memory aided nature of vision,clearly hints at need of having inherent class specific representation at intermediate levels of CNN. We therefore propose, deep fusion architecture which not only combines class-conditional representations but also introduces scale invariance by combining it via inception[18] like module.

2. Related Work

The inception of saliency based algorithms was spearheaded by Itti et al.[10],which used dynamic neural networks and multiscale features for generating image saliency maps. They used rudimentary image features such as color,orientation and contrast in order to select and conspicuous locations. In extension to Itti Et Al, Harel Et Al's approach[7] tends to highlight conspicuity based on redistribution/normalization of pixel scores which are calculated by construction of activation map. These maps, which are similar in construction to Itti's model[10], are processed using Markov chains, to obtain activation values at equilibrium.

On similar lines of Feature maps based segregation,some recent works include the one by Guabin Li Et Al [14],which uses multiscale feature extraction by having multiple input crops of salient region, each having different contextual information. It perform multilevel region decomposition of image, followed by weighted reconstruction of the heatmap. However,we can also replace these manually crafted feature maps by appropriate CNN. We tend to exploit this inherent capability of CNN by using inception module in order to achieve some level of scale invariance. We also create customized feature pyramid which serves dual purpose of feature augmentation for enhanced final saliency calculation and a pathway for effective back-propagation to tackle vanishing gradients.Further, these activation maps are converted to binary maps using dense CRF which takes care of context aware segmentation. Similar work of using unsupervised feature map refinement has also been conducted by Huaizu Jiang et al[?] which uses context based unsupervised approach for salient object segmentation. It tries to discriminate superpixel regions based on histogram similarity and other parameters such as ratio of areas of current region with neighbour, etc. These scores are calculated using superpixel segmentation at different scales, and also consist of smoothness term. This term is responsible for segmentation boundary to be aligned with computed closed contour, using graph flow algorithms. Use of Dense CRF in our algorithm takes care of context aware smoothness of the segmentation map. Huaizu Jiang et al[?] also incorporates edge detection based shape prior which is combined based on overlap and average scores. Such object category/shape biased priors are quite plausible in visual apparatus of primates.This feature is entailed in our CNN by choosing pretrained class conditioned model. It allows CNN to incorporate this object/shape bias in the final representation.



3. Proposed Approach

As discussed earlier, the approach consists of designing CNN model for supervised training followed by unsupervised refinement of saliency map via Dense CRF[13].

3.1. Fusion Architecture

We use pretrained model obtained from deeplab[3] which gives 21 binary output maps corresponding to PASCAL VOC 2012 dataset. These class conditioned representations can be though of as representations of object assuming it belongs to that category.

$$F_i(x; w) \approx P_w(x|c = i) \tag{1}$$

In Equation 1, $F_i()$ represents ith slice of overall CNN response with $i \in [0, 20]$ corrosponding to Pascal VOC 2012 classes. As shown in Figure 1, main branch of the net consists of series of interleaved convolution and pooling layers. The auxilary branches stemming out of response of pooling layers in main branches are passed through inception layer. These responses are then concatenated and combined via a final inception module resulting into final saliency map. Each inception module mainly concatenates responses of filters with different support with necessary padding. Larger the support/size of filter in inception module, smaller is its contribution to final depth because of huge memory constrain encounter for the filters with larger support.

Inception module used in our CNN model is variant of the one present in googLeNet[18] as illustrated in figure 1. The overall depth of our inception module is very less as it's not a part of main branch where depth of inception module's response makes a difference. However, the underlying principle of considering contributions from all scales is same as mentioned by Szegedy Et Al[18].We use this to obtain multiscale feature pyramid depicted in Figure 1 as response of 'final concat' layer, with size (54x65x65). This concatenation of intermediate responses help in combining multilevel representation. It also gives a plausible solution to a higly probable issue of vanishing gradients while backpropagation as sensed by Szegedy Et Al[18]. This is the sole reason behind GoogLeNet having three pathways for backpropagation.Provision for such alternate pathway is one of the key reasons for taking direct contribution from these auxilary branches as it allows an alternate pathway for backpropagation via inception modules.Use of pretrained VGG Net consists of Class conditioned feature maps which were created by taking euclidean difference between response of 'FC8-VOC12' shown in figure 1 with size 21x65x65, and its ground truth. We assert that these intermediate representations are decentralized and diversified by introducing initial class biased model. The final class conditioned map is also fused and concatenated through inception like module[18]. The learning rates for fusion layer is set very high followed by rest of inception modules. This ordering, allows us to manipulate and force learning ,through different layers.
The obtained saliency map is thereby used to initialize unary potential of Dense CRF[13]. The pairwise cost is based on local features as discussed in the previous section.

3.2. Dense CRF

Conditional Random Fields come into picture, when we want to fuse the posterior probabilities obtained via CNNs. This is done by assigning unary potentials as these probabilities. Dense CRF[13] not only accounts for pairwise interactions, but also models interaction of a pixel with all other pixels, in a weighted manner . If Φ_u and Φ_p are unary Potential and pairwise potential, then we try to maximize negative of Gibbs Energy Given by :

$$E_G = \sum_{\forall \vec{x} \in V_G} \Phi_u + \sum_{\forall (\vec{x_1}, \cdots \vec{x_n}) \in C_G} \Phi_p(x_1 \cdots x_n) \tag{2}$$

In,equation 2, C_G,V_G stand for set of cliques and set of vertices in graph G.Pairwise potential of the model is given by :

$$\Phi_p(\vec{x_i}, \vec{x_j}) = \mu(\vec{x_i}, \vec{x_j}) \sum_{(\vec{x_i}, \vec{x_j}) \in E_G} w(\vec{x_i}, \vec{x_j}) K(\vec{x_i}, \vec{x_j}) \tag{3}$$

Here, $\mu(\vec{x_i}, \vec{x_j})$ is the label compatibility function and $K(.)$ is gaussian kernel given by $exp(-\frac{1}{2}(f(x_i) - f(x_j))^2 \Lambda (f(x_i) - f(x_j))^2)$. Here Λ is positive definite precision matrix.
In general, the objective function to be minimized in Dense CRF is given as follows :

$$w_a exp \overbrace{\left(-\frac{|p_i - p_j|^2}{2\sigma_{ap}^2} - \frac{|I_i - I_j|^2}{2\sigma_{ai}^2} \right)}^{\text{appearance kernel}} + w_s\, exp \overbrace{\left(-\frac{|p_i - p_j|^2}{2\sigma_{sp}^2} \right)}^{\text{smoothness kernel}} \tag{4}$$

Above Equation is for kernel $K(f_i, f_j)$ over feature vectors(in our case it is local weighted histogram). Here p_i, p_j indicates position vectors for a given graph, and I_i, I_j indicate the value of Image/Function for the pair. Note that auxiliary parameters are not static, but approximated each time a new image is passed for segmentation.

4. Experimental Setup

We use MSRA10K dataset for finetuning and validation of MSRCOCO and Pascal pretrained VGG-16 Net[3]. This model has converted inner product layers to convolutonal layers apriory. The images present in MSRA1k are used for validation, reserving the rest 9K images for finetuning of pretrained VGG-16 Net[3]. We also use CSSD Dataset consisting of structurally complex images for validation. Both the validation data have binary ground truth maps

against which the evaluation is performed
We use Mean Absolute Error(MAE) over Dense CRF refined results as the primary metric of evaluation. Precision and Recall measure over saliency maps exuded from CNN forms the secondary metric of evaluation of our approach. The evaluation results are shown in table **??** and are compared with existing approaches such as Geodesic Saliency(GS),Saliency Filter(SF),Saliency Optimization(SO) and Manifold Ranking(MR).

4.1. Mean Absolute Error

It measures the average pixelwise similarity of ground truth with the predicted binary map as shown in equation 5

$$MAE = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{WH}\sum_{j=1}^{W}\sum_{k=1}^{H}|gt^{(i)}(j,k)-o^{(i)}(j,k)| \quad (5)$$

In equation 5, W and H stand for width and height of the image. C represents total count of groundtruth (gt) and predicted map (o).

Comparison of Mean Average Error

Dataset	GF[**?**]	SF[**?**]	MR[**?**]	SO[**?**]	**Proposed**
MSRA1K	0.109	0.129	0.085	0.068	**0.0628**
CSSD	0.109	0.129	0.085	0.068	**0.0628**

4.2. Precision and Recall

Precision is the ratio of cumulative correctly assigned salient pixels to total salient pixels in image in all images.However recall represents the ratio of total correctly assigned pixel to total number of pixel, irrespective of it being salient or not. The saliency maps resulting from CNN,are thresholded over different values, in order to observe relationship between precision and recall. We also compute F measure given by :

$$F_\beta = \frac{(1+\beta^2)Precision.Recall}{\beta^2 * Precision + Recall} \quad (6)$$

In equation 6, β^2 is set to 0.3 in order to lay more emphasis on precision then recall.

5. Observations

As we can observe in figure**??**, the response of simple fusion has extra regions fails to encorporate region localization as effectlively as inception aided experiment.

6. Conclusion

6.1. Experimental Setup

References

[1] E. Averbach and A. S. Coriell. Short-term memory in vision. *Bell System Technical Journal*, 40(1):309–328, 1961. 1

[2] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 1

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1, 3

[4] J. H. Dholakiya, R. K. Sarvadevabhatla, and R. V. Babu. Expresso: A user-friendly gui for designing, training and using convolutional neural networks. *arXiv preprint arXiv:1505.06605*, 2015. 2

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. 1

[6] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391(6666):481–484, 1998. 1

[7] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006. 1, 2

[8] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 1

[9] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10):1304–1318, 2004. 1

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998. 1, 2

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 2

[12] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1072–1080, 2015. 1

[13] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*, 2012. 1, 3

[14] G. Li and Y. Yu. Visual saliency based on multiscale deep features. *arXiv preprint arXiv:1503.08663*, 2015. 2

[15] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381. ACM, 2003. 1

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein,

et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014. 1

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 1, 2, 3

[19] S. Zeki, J. Watson, C. Lueck, K. J. Friston, C. Kennard, and R. Frackowiak. A direct demonstration of functional specialization in human visual cortex. *The Journal of neuroscience*, 11(3):641–649, 1991. 1