iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5118

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Literature Review on Object Detection with Backbone Network

Rakshitha S G[1], Revathi S A[2]
[1]*Student, Department of Computer Science & Engineering, R V College of Engineering, Karnataka, India*
[2]*Affiliated to VT University, Belagavi, Karnataka, India*

Abstract: ***The development of Salient Object Detection is the realization of all stirring objects in the image or videos etc. It is useful in various applications of computer vision such as in video analysis, face recognition, understanding of images, and tracking of objects in games and it is useful in medical-related applications. In Computer Neural Network (CNN) object detection can be done using various approaches, this paper will review how object detection can be achieved using backbone networks. The backbone network is based on deep-lab architecture, it is also called as the feature extracting network. It takes input as an image and then it extracts the feature of the image and then produces the feature map, that feature map will be used by the remaining networks in-order to detect the object and produces the output as the detected object. A decade ago, object detection was done on CNN using the conventional method, it was less efficient compared to new methods evolved in CNN. The Encoder-Decoder architecture on CNN has appeared to be more efficient, powerful and it is helpful in salient object detection.***
Keywords: Computer Neural Network (CNN), Residual Neural Network (RESNET), Visual Geometry Group(VGG), Backbone Enhanced Network (BENET).

I. INTRODUCTION

Salient object detection can be done in two ways that are, by segmenting the object out from the background of image or video, etc and detecting those hidden objects from the background. It is a serious challenge in numerous actual scenes, because objects may have different colors, texture also properties as background. Newly Encoder-Decoder architecture is used in CNN for object detection.

In this architecture the Encoder will be using backbone networks such as Residual Neural Network (ResNet) and Visual Geometry Group (VGG) for getting feature maps with high-level multiscale information about the salient objects. Here parameters will be pre-trained with the ImageNet database. These parameters will be re-trained with a dataset of the salient object. The boundary detected by the re-trained network will be added with a clearer and complete boundary line through class activation mapping (CAM), compared to pre-trained parameters of the network.

The dual backbone network is adopted in the encoder to maintain stability in generalizing the object features and precision. In this Encoder the dual backbone will be having two networks each network will be pre-trained with the Image Net dataset to produce efficient feature maps.

CNN has different layers each layer is called a feature map. In these two networks one network will be having the pre-trained datasets and another will be retrained this will hold the new re-trained dataset this will make use of both pre-trained and re-trained datasets so that both features can be analyzed efficiently to come up with more useful and effective feature map. This feature map will be helpful to identify the salient objects with including there even small details like color, size, etc. This will help to create the proper boundary of that object so that object can be easily identified. RS2Net connection module will be added in the decoder to combine the data captured from both the networks and multi-scale features will be added to a decoder which provides decoder to multi-scale training to produce a weighted multi-scaled feature map.

II. METHODOLOGY

Salient object detection using Backbone Enhanced Network (BENET) is like Feature Pyramid Networks for object detection. Backbone Enhanced Network includes components such as encoder, decoder, connection module in U-shaped structure. The dual backbone network is adopted in the encoder to maintain stability in generalizing the object features and precision.

The connection module is used to connect two networks dataset and sends it to decoder, the decoder will be having a multiscale feedback module (WMFM) to generate the efficient feature map.

BENET is as shown in the below figure,

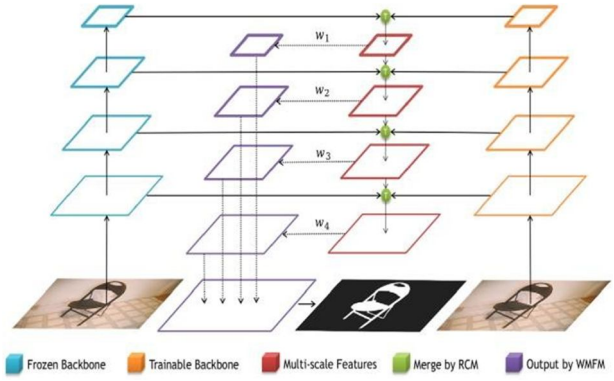


Fig.1. BENet Framework. A different form of feature maps is indicated in distinct colors.

A. Dual Backbone Network

The encoder in feature pyramid network will be of the classic backbone network, it uses the bottom-up structure to extract features into a multiscale featured map. In feature pyramid network, the encoder is usually a classic backbone network with a bottom-up structure. And multi-scale feature maps are extracted with the function shown in Equation,

$$Y_j = \sum_{i=0}^{C_{in}-1} W_j(k,p,s,i) \circ X_i + B_j$$

where Xi is the ith input channel of the feature map, output in feature maps are indicated as Yj and Wj, Bj, the derivation of function convolution kernel of the output channel jth it shows the operation of convolution, the number of input channel indicated by Cin, Kernel scale indicated by k, size of padding indicated by p, step size indicated by s. The popular feature pyramid in backbone networks is ResNet-50 and VGG-16 networks. And ResNet-50, VGG both have information about pre-trained parameters which was generated using ImageNet. which can extract features of almost all objects of classes that were in ImageNet. So that many methods make-use of the pre-trained parameters without having to retrain and use them directly for getting feature maps. Here they use CAM, from the result of this mapping people can analyze the pre-trained parameters using CNN to extract features from different image regions to identify the object type its regional locality in the image, etc.
tells us that pre-trained parameters in the network are used for getting regional locality information of the object. That means the pre-trained network mainly focuses on region features for object recognition. In our approach the parameter pre-trained will remain the same but new features will be trained on other network parameters so that pre-trained features and new features trained parameters both retained. This makes avoiding the over-fitting of the dataset on which the parameters are trained. It is useful in-order to efficient object mapping or identification it avoids adjustment of training images. But sometimes it may perform poorly because the retrained dataset may be more different from a pre-trained dataset using ImageNet.

B. Res2Net based Connection Module

The feature maps captured by the two-backbone networks in Dual backbone network (DBN) is to achieve information fusion and Res2Net based Connection Module (RCM) are used to process the information, as in below fig,

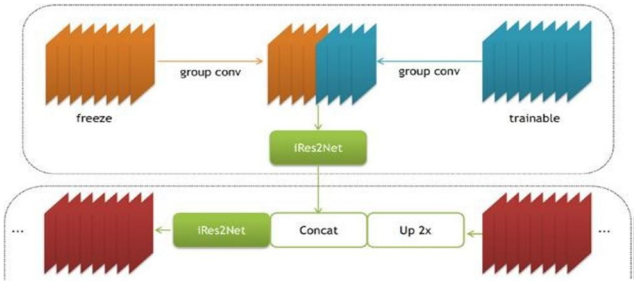


Fig.2. Connection Module based on Res2Net.

The DBN retrieves two feature maps they are indicated as blue and orange feature maps; these feature maps are joined using group convolution and iRes2Net is used to process the maps. The iRes2Net output is then joined with low-resolution feature maps at a lower level. To reduce the number of channels in feature map the DBNs two network extracted feature map will be useful. To extract features from every channel that are present in network Group convolution will be useful. In-order to fetch even more meaningful data from images, iRes2Net is included, the output generated from this iRes2Net will be given to the decoder. The iRes2Net is adopted instead of Res2Net, because In-order to increase the possibility of fusion feature between various channels and the add operation feature is replaced by the concatenate operation.

C. *Weighted multi-scale Feedback Module*

For up-sampling in decoder the bilinear interpolation is used, the feature of up-sampled will be concatenated with the RCM output. Then iRes2Net is used to process the concatenated feature maps. In this way, the decoder combines semantic information and puts that information in a high-level and for low-level spatial details will be stored. Usually the salient region or layout will be blurry because if there is any lack of information in generated maps. Identifying what features are salient the high-level feature maps will be helpful. The high-level feature maps help find what is salient, but sometimes the salient region is blurry due to the lack of detailed information in those maps. Low-level feature maps will be having object detailed information, but it will be more difficult to identify the saliency of these low-level feature maps. But one solution is to maintain the balance between feature maps using iRes2Net and integration of Multi-level feature maps to extract details accurately.

Our generated feature map will increase the network depth with RCM and iRes2Net in the decoder. But this will eventually produce problems related to backpropagation. That is, it cannot effectively transfer the loss of a network from layers present on a network called the deep network to the shallow layer. To get-ride of this problem, the solution is to use multi-scale direct feedback and decrease the path from upper-level layers such as supervised layers to low- level layers called shallow layers. In each layer's supervision will be performed which are present in a decoder. This supervision will adjust the contribution of features in each layer and adjust the contribution and reverse the loss, if any loss occurs it sends it to the shallow layer. In different scenarios the distance between the low-level feature maps to ground truth mask will be smaller compared to the distance between the high-level layer to low-level layer feature maps. Calculate the weight automatically and decide to use the convolution parameter, then the weight of lower resolution layer feature maps will be set to zero after so many rounds of training. So, the loss of a lower feature map layer cannot be sent to the shallow layer. To solve this problem each feature map layers will be assigned with fixed weight generated by the scale and number of channels on the network. The output of weighted feature map Fj calculated at level j using the below-mentioned equation.

$$F_j(X_j) = X_j \times \frac{w_j \times h_j \times c_j}{\sum_{i=0}^{N} w_i \times h_i \times c_i}$$

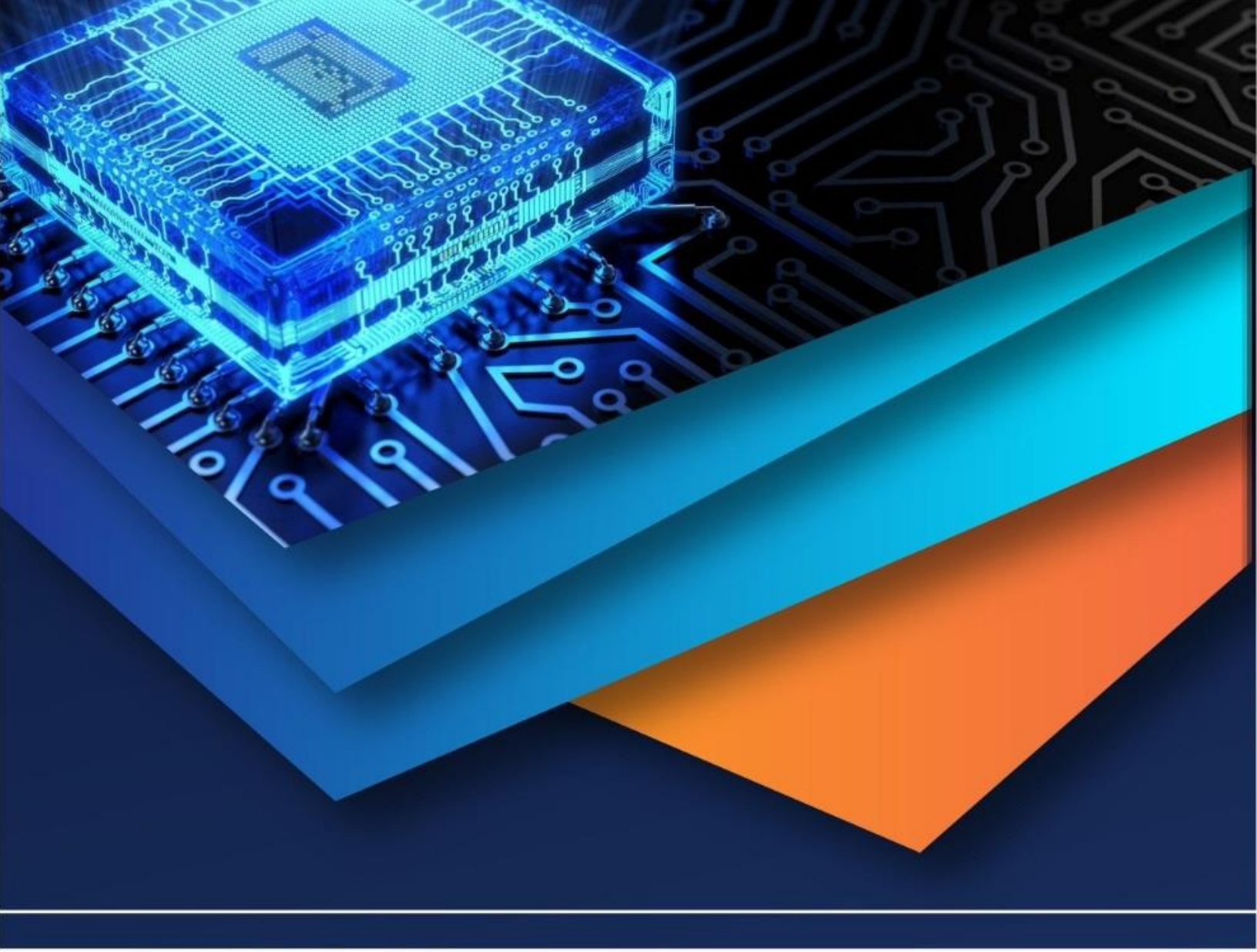
D. *DetNAS Pipeline*

DetNAS pipeline is used to find optimal backbones in object detection without loss of generality. This can be included as an additional feature in order to reduce the generality DetNAS pipeline has three stages:-

1) *Supernet pre-training:* In the backbone network, the pre-training dataset is a basic task. In the pre-training dataset discrete space search will be identified and made it to one space. But in supernet training it trains the data pathwise. This supernet will examine the performance of the network by iterations. Paths will be iterated, and these paths will be sent to the supernet graph. The graph will not perform operations on the updated weight, but it will take the paths sent by the supernet.
2) *Supernet fine-Tuning:* Supernet fine-tuning will include the metrics and detection head for fine-tuning the supernet and it uses normalization methods while fine-tuning the dataset. In this BN parameters will be fixed as a pre-training statistic. Fixing the BN parameters may reduce accuracy. So SyncBN was introduced to balance the batch size and accuracy.
3) *Search on Supernet:* On trained supernet search will be done. In this search the paths will be picked and then it will be evaluated. Different child networks will be evaluated and sampled. Each sampled data should be dependent. In this paper, discussion was held on how to use a dual backbone network with Encoder- Decoder Architecture for Salient Object Detection. To combine maps generated from two networks, the Res2Net connection module is used and the Feature pyramids helps several scales in detecting the objects. The multi-scale weighted feature is useful in final feature map generation and it can be trained to achieve multi-scale supervision. In Decoder bilinear interpolation, to separate segmentation features at a high level and special details at a low level, iRes2Net module is used and discussed about DetNet pipeline with three stages.

REFERENCES

[1] A. Borji, M.M. Cheng, “Salient Object Detection”, 2014, Comput. Vis. Media.
[2] 1. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real- time object detection. Computer Vision Pattern Recognition Proceedings IEEE Conference . (2016)
[3] Zhu, M., Chen, B., Howard ., A.G,Kalenichenko, D, Efficient convolutional neural networks for object detection applications.(2017)
[4] Nathan Tsoi, Ian Reid, and Silvio, Hamid Rezatofighi, Savarese. A metric loss for bounding regression box, 2019.
[5] Jingdong Wang, Ke Sun, Bin Xiao, Dong Liu . Deep high-resolution learning estimation for human pose estimation 2019.
[6] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring, 2019.
[7] Xiaoliang Dai, Bichen Wu, Yanghan Wang, Peizhao Zhang, Yiming Wu, Fei Sun. Fbnet: Hardware-aware efficient Design differentiable neural architecture search, 2019.
[8] Hehui Zheng, Sirui Xie, Liang Lin ,Chunxiao Liu. stochastic neural architecture search. 2019.
[9] Tong Yang, Zeming Li, , Xiangyu Zhang, Jian Sun, Wenqiang Zhang. Learning to detect objects customized anchors. 2018.
[10] Xizhou Zhu, Stephen Lin, Han Hu and Jifeng Dai. More deformable,better results. 2018.
[11] Barret Zoph Quoc V. Le. Neural architecture search reinforcement learning. 2016.
[12] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Learningtransferable architectures for scalable image recognition.2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY