The multi-scale channel attention full connection network for edge detection

Meng Li¹, Wei Wu²

¹Department of Information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430070, China
²Department of Information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430070, China

*Corresponding author’s e-mail: wuwei@whut.edu.cn

Abstract. Edge detection algorithms based on deep learning have made great progress due to its super feature extraction and presentation ability in recent years. However, existing edge detection algorithms encounter some problems that image features extracted by networks are not sufficient, and the generated edges are fuzzy with large gap compared to the ground truth (GT). To solve the abovementioned problems, we have proposed a multi-scale channel attention full connection network model for edge detection (MAED) in this work. The proposed network consists of three parts: multi-level feature extraction part, multi-scale deep supervise part and fusion part. In addition, by introducing the "attention" mechanism into the networks, underlying features of the input image can be fully extracted and the training convergence speed can be accelerated simultaneously. To evaluate the performance of the proposed algorithm, we conducted a comprehensive study on some commonly-used datasets. The final experimental results demonstrated that our network achieves excellent performance based on both qualitative and quantitative analysis.

1. Introduction

Edge detection, aiming at extracting the most basic and important features in images, has always been playing an important part in the field of image processing and pattern recognition. The edge maps obtained by image edge detection present the basis information for image analysis, target detection, target recognition, image segmentation, etc. To sum up, how computer reacts to the real world depends on the accuracy and reliability of edge detection results in practice.

The traditional edge detection operators are constructed based on the property that grey level of a small neighborhood of the edge pixel changes dramatically. The operator enables extraction of the corresponding edges in the image by convolution operation. After the introduction of Sobel operator, Canny operator, Laplace operator and other methods have been proposed one after another, which have been used ever since, but the classical edge detection is much susceptible on the artificial threshold, making it improper to be used in the general case. Fortunately, with the improvement of computing power, convolutional neural networks have prevailed in the field of image processing with their super feature extraction ability. Many studies have found that edge detection algorithms based on deep learning are not only accurate, but also superior to traditional algorithms in detection speed. In this condition, edge detection algorithms based on deep learning become the mainstream methods in edge detection tasks gradually.
Since Ganin[1] et al. first tried to introduce simple combination of CNNs and nearest neighbor search, CNNs have brought prosperity to edge detection. The designs and training strategies of various network architectures continuously improve the performance of edge detection. The end-to-end edge detection model Holistically-nested edge detection[2] proposed, embodies the idea of multi-scale features for the first time. In addition, the residual network also has been applied to edge detection problem recently. With network depth increasing, the better the feature expression ability. Inspired by these excellent works, a network based on residual block and channel attention mechanism is proposed to capture more global features and fully fuse multi-scale features in this paper. This network combines the advantages of previous work, and improves edge detection accuracy and clarity.

2. Related Work
It is not difficult to find that the features are more robust, the result is much better. However, the predecessor's strategy is still based on local features of picture, Therefore, the actual performance is far from that of humans.

The DeepEdge[3] proposed in 2015 extended this work. Firstly, they used traditional Canny operator to obtain candidate contour points, then introduced a multi-scale technique to divide these points into four different scale patches. In the design of network, it was divided into two tasks: classification and regression, which further improved the robustness of model. The above prepared contour points were fed into two branches, and the probability of each candidate point was added finally. However, the network depth of this method is still shallow, and some high-level information is not learned. The VGGNet proves that there is a close relationship between network depth and performance. It is a light structure and has strong generalization ability, which can be easily extended to different datasets. HED used the VGG-16 trained by high-level semantic information as the backbone, initialized the network weight by migration learning, and obtained the high-level information of whole image. Inspired by HED, RCF[4] used VGG-16’s all conv layers to fuse multi-scale information to improve detection performance.

Hierarchical networks can obtain multi-level feature maps from inner layer of CNNs, while attention mechanism further enhances its feature representation ability. Since AMH-Net[5] just considered the human-visual based understanding, the attention gating condition random field (AG-CRF) in AMH-Net enabled to learn robust feature representation on each scale by exploring and utilizing the information provided by other scales, which could achieve better results. He et al. believed that supervise learning applied to all conv layers with the same scale in previous models is defective. On the contrary, different layers should be supervised with specific scale. Scale Enhancement Module (SEM)[5] was introduced to generate multi-scale features using dilated convolution without fusion of multi-scale feature.

3. Proposed Methods
In this chapter, we will introduce the proposed network in detail. As shown in figure 1, our network consists of three parts: multi-level feature extraction backbone part, multi-scale supervise part, and feature fusion part.

3.1. Multi-level feature extraction backbone part
The main backbone is inspired by modified ResNet50[6]. Feature extraction part only uses its first 5 layers and final full connection layer is removed. Each branch is connected to the side of each stage for deep supervise learning. Inspired by the notion that network depth affects feature extraction ability, we use residual structures in backbone. However, unlike classical ResNet50, an effective cross-channel attention module (ECABottleneck)[7] is introduced in our paper, which only adds a small number of parameters but with better performance gains. Not only can the attention module emphasize useful features and suppress useless features through learning, but also the balance between network performance and time complexity is considered. First two layers of backbone are used to extract the low-level features by using traditional bottleneck; the last three layers use ECABottleneck to avoid
dimensionality reduction and appropriate cross-channel information interaction at the same time, which can extract more global features.

3.2. Multi-scale supervise part
As shown in the middle part of figure 1, our multi-scale supervise block uses similar deep supervise technology in RCF. First two layers use 1*1-21 conv layer for feature compression and the rear three layers use 2*(l-2) 2*(l-2)2* 2 1  conv layer alike. Traditional bilinear interpolation method, where a large number of regions are filled with 0 directly, will lead to interference factors and lose some useful information. Inspired by the work in field of super-resolution[8], this paper uses the upsampling method of pixel shuffling to restore the larger feature map and the single pixels on the multi-channel form the unit on the feature. Pixels on each feature are equivalent to subpixels on new extracted feature with more real information. The results show that small curved edge, such as hand, enhance the correlation of information; the information utilization rate is higher and edge detection performance is improved.

3.3. Fusion part
The fusion block is responsible for the fusion of five cascaded outputs. In HED and RCF, there are only five stages connected in equal proportions, after extracting the spliced map with kernel size 1*1. Such fusion is much simple and crude to take into account the importance of different levels of information. Therefore, we also improve here, hoping to achieve local cross-channel information interaction, that is, how many adjacent channels near the channel are involved in the attention prediction of this channel where use adaptive kernel size to determine this coverage. After such sufficient processing, a 1*1-1 conv layer is used for feature compression for final high-quality edge probability map.

3.4. The proposed loss function
In our model, edge detection is regarded as a pixel-level binary classification problem, which the pixel points are divided into edge points and non-edge points. Because the number of pixels as edge points in the image is far less than non-edge points, the weighted cross entropy is used as the classification loss function of each pixel.

On the basis, ground truth of dataset is marked by different people, there is inevitably some cognitive errors. So for each image, we average all the ground truth to generate an edge probability map, which range from 0 to 1. Here, 0 means that no annotator labels on this pixel, and 1 means that all annotators label on this pixel. The pixels with edge probability higher than $\eta$ are considered as edge pixels, and
pixels with edge probability equal to 0 are regarded as non-edge pixels. If a pixel is labelled fewer than \( \eta \) persons, the pixel is considered as a controversial pixel and should be abandoned. The label equation as below.

\[
\text{label}(X_i; W) = \begin{cases} 
\alpha \times \log(1 - P(X_i; W)) & \text{if } y = 0 \\
0 & \text{if } 0 < y \leq \eta \\
\beta \times \log P(X_i; W) & \text{if otherwise}
\end{cases}
\]

(1)

in which

\[
\alpha = \lambda \times \frac{Y^+}{Y^+ + Y^-}, \quad \beta = \frac{Y^-}{Y^+ + Y^-}
\]

(2)

\( Y^+ \) and \( Y^- \) represent the number of positive samples and the number of negative samples, separately. The hyper-parameter \( \lambda \) is used to balance between positive and negative samples. \( X_i \) denotes the feature vector of convolution neural network at the pixel \( i \) and \( y_i \) is ground truth edge probability at pixel \( i \).

Finally, our loss function is that weighted sum of loss of fusion layer and loss of each stage layer.

In the course of the experiment, we find that the importance of each stage loss is different from the fusion loss, because each stage loss only contains local information loss, which is an auxiliary function to the whole network. The fusion loss is the final output of the model and plays a decisive role. Through repeated debugging, when \( \gamma \) set to 1.3, it can explicitly improve performance, as shown below.

\[
L(W) = \sum_{i=1}^{K} \sum_{k=1}^{K} \text{label}(X_i^k; W) + \varphi \times \text{label}(X_i^{\text{fuse}}; W)
\]

(3)

Where \( X_i^k \) represents excitation value of the image output from layer \( k \) at the pixel \( i \), \( X_i^{\text{fuse}} \) is excitation value of the image output from the fusion module at the pixel \( i \), |I| represents the total pixel of the image, \( K \) is number of stage in Resnet50 (\( K = 5 \)), \( \varphi \) is the hyper parameter to balance the two part losses.

4. Experiments

4.1. Datasets and metrics

A popular open source dataset Berkeley Segmentation Dataset (BSDS500)[9] is used to train our network, but original BSDS500 dataset has only 500 images, including 200 train images, 100 verification images and 200 test images. Obviously, the number of train pictures is too small, which will lead to overfitting of the model. As in HED[2], we extend the data of 300 pictures of train dataset and verification dataset, including rotation, flip, scale scaling, etc. to obtain 28800 images. After training, we compare the model performance in some standard datasets: BSDS300[10], BSDS500, NYUD[11]. We have used evaluation indicators from the pioneering work: optimal dataset scale (ODS) which employs a fixed threshold for all images in the dataset, optimal image scale (OIS) which selects an optimal threshold for each image, F-score of the ODS and OIS will be considered, and AP.

4.2. Training Details

The development framework used in this experiment is pytorch 1.4.0 which is well-known in this community. Recent literature has demonstrated that pre-trained fine-tuning depth neural networks on general image classification tasks are useful for edge detection task. So we use the pre-trained model Resnet50 on the ImageNet to initialize our network.
In formal training, global learning rate sets 0.01, then polynomial attenuation strategy is applied to dynamically adjust learning rate. From initial learning rate to lowest learning rate, after reaching lowest learning rate, learning rate is increased to a certain value, then reduced to lowest learning rate, so that the model cannot fall into local optimal solution. Using SGD as optimizer, the momentum sets 0.9 and weight decay is set to $5 \times 10^{-4}$.

4.3. Result of Test

In order to improve the quality of edges, we use the image pyramid technology in test period. The original images are adjusted to 0.5 and 1.5, later three scales are fed into our single-scale detector, respectively. The obtained edge map are restored to the original size by bilinear interpolation, and final map is obtained on average. Although speed is sacrificed, the results show this is useful for accuracy.

After generated maps are treated with NMS, they are measured by evaluation script, We compare the ODS, OIS and AP of our model with HED, RCF, CED[12] and SED[13] on some classical datasets as is shown in table 1. From quantitative analysis of table 1, our performance of network is much higher than RCF. Compared with CED and SED, our network shows better performance on BSDS500.

| Dataset  | HED | RCF | CED | SED | ours |
|----------|-----|-----|-----|-----|------|
| BSDS300  | ODS | 0.690 | 0.708 | -   | 0.690 | 0.708 |
|          | OIS | 0.701 | 0.712 | -   | 0.710 | 0.708 |
|          | AP  | 0.700 | -    | -   | 0.710 | 0.590 |
| BSDS500  | ODS | 0.790 | 0.806 | 0.803 | 0.710 | 0.810 |
|          | OIS | 0.808 | 0.823 | 0.820 | 0.740 | 0.837 |
|          | AP  | 0.811 | -    | 0.871 | 0.740 | 0.842 |
| NYUD     | ODS | 0.720 | 0.743 | 0.726 | -    | 0.730 |
|          | OIS | 0.761 | 0.757 | 0.750 | -    | 0.761 |
|          | AP  | 0.786 | -    | 0.778 | -    | 0.780 |

We also show some visualization images generated by edge detection algorithms in figure 2. Because the performance of RCF model is best, in visual comparison, we only choose RCF and MAED for comparison. For the picture "Plane", we can see that our "edge plane" and GT are closer, introducing less noise. Edge image generated by our model is significantly better than RCF. The edge lines generated by RCF are much vague in detail, while we have more details than RCF. For the picture "Boat", we can see that small line like the distant sail still generate better than RCF.

5. Conclusion

In this paper, we propose an effective multi-scale channel attention model for edge detection task. The architecture consists of three parts: (1) multi-level feature extraction backbone part, (2) multi-scale supervising part, and (3) fusion part. In feature extraction part, we use ECA Bottleneck, which enhances characterization ability of backbone without combining additional spatial attention mechanism. In supervise part, the features extracted by the previous network is supervised separately and upsampled by sub-pixel convolution, so that the edge lines are continuous and clear. Finally, the skip connection is used to fuse all features obtained by the network. The experiment proves that our method is effective and can produce higher quality edges, which makes it promising for visual tasks such as instance segmentation, significant target detection.
Figure 2. Comparison of original image, GT, RCF and our image (from left to right).

Acknowledgments
Thanks to my colleagues and mentor for their assistance, as well as the financial support from the National Natural Science Foundation of China: Research on the visual perception enhancement method of unmanned surface ships in severe navigable environment.

References
[1] Y. Ganin and V. Lempitsky, $ N^4 $-Fields: Neural Network Nearest Neighbor Fields for Image Transforms. 2014.
[2] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," International Journal of Computer Vision, vol. 125, no. 1, pp. 3-18, 2017/12/01 2017.
[3] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multi-scale bifurcated deep network for top-down contour detection," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4380-4389.
[4] Y. Liu et al., "Richer Convolutional Features for Edge Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, pp. 1939-1946, 2019.
[5] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe, "Learning Deep Structured Multi-Scale Features using Attention-Gated CRFs for Contour Prediction," 01/01 2018.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.
[7] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020, pp. 11531-11539.
[8] W. Shi et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874-1883.
[9] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour Detection and Hierarchical Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 898-916, 2011.
[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, 2001, vol. 2, pp. 416-423 vol.2.
[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in Computer Vision – ECCV 2012, Berlin, Heidelberg, 2012, pp. 746-760: Springer Berlin Heidelberg.
[12] Y. Wang, X. Zhao, and K. Huang, "Deep Crisp Boundaries," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1724-1732.
[13] A. Akbarinia and C. A. Párraga, "Feedback and Surround Modulated Boundary Detection," *International Journal of Computer Vision*, vol. 126, 07/27 2017.