Comparison of Backbones for Semantic Segmentation Network

Comparison of Backbones for Semantic Segmentation Network

Rongyu Zhang[1], Lixuan Du[1], Qi Xiao[2] and Jiaming Liu[3,*]

[1]International School, Beijing University of Post and Telecommunication, BEIJING, 100876 CHINA
[2] International School of business&Fiance, Sun Yat-Sen University, GUANGZHOU, 510275 CHINA
[3]Pattern Recognition and Intelligent System Lab, Beijing University of Posts and Telecommunications, BEIJING, 100876 CHINA

* Corresponding author email：liujiaming@bupt.edu.cn

Abstract. As for the classification network that is constantly emerging with each passing day, different classification network as the backbone of the semantic segmentation network may show different performance. This paper selected the road extraction data set of CVPR DeepGlobe, and compared the performance differences of VGG-16 as the backbone of Unet, ResNet34, ResNet101 and Xception as the backbone of AD-LinkNet. When VGG-16 is used as the backbone of the semantic segmentation network, it performs better in the face of long and wide road extraction. As the backbone of the semantic segmentation network, ResNet has a higher ability to extract small roads. When Xception is used as the backbone of the semantic segmentation network, it not only retains the characteristics of ResNet34, but also can effectively deal with the complex situation of extracting target covered by occlusions.

1. Introduction

With the continuous development of deep learning technology, researchers began to apply it in various fields, including the field of computer vision. Before this, people could only use Texton Forest or Random Forest as the semantic segmentation classifiers. However, nowadays, the classification networks such as AlexNet[1] and VGG[3] brought by the emergence of convolutional neural network (CNN) have greatly improved the semantic segmentation ability of computers. In this paper, we will compare the advantages and disadvantages of various classical basic classification networks as backbones through experiments, but not including some networks with special functions, such as SSD for detection or FCN for special segmentation.

2. Background

2.1. Classification network method at the present stage

We use ImageNet classification errors as a standard to review the classic networks in recent years [2].

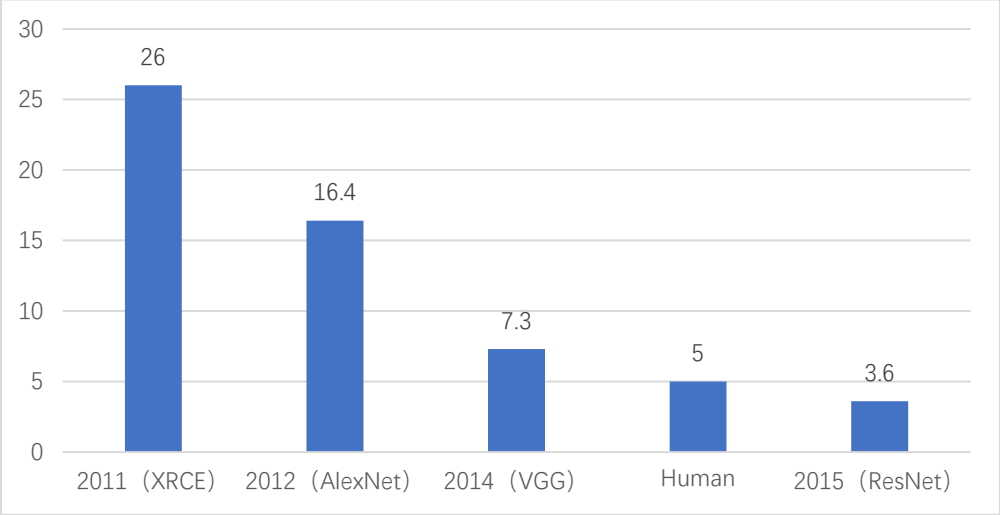

Figure 1. ImageNet classification errors

In 2012, AlexNet introduced CNN, which significantly widened and deepened the network, and replaced the average pool with the maximum pool, making the network classification error significantly decreased compared with the previous year's XRCE, which is a qualitative leap for deep learning.

In 2014, VGG reduced the classification error to less than 10 and greatly increased the number of network layers to 16-19. Compared with AlexNet, VGG uses smaller convolution kernel and pooling kernel, which increases network depth and reduces parameters. Meanwhile, since the Local Response Normalization layer is not effective, VGG removes the LRN layer [3].

ResNet in 2015 took deep learning to the next level, surpassing humans (5) with 3.6 classification errors and extending the network to a depth of 1202 layers. The key to ResNet's success lies in the adoption of the core idea of Identity Shortcut Connection [4], which solves the problem of "As the network deepens, accuracy does not decline".

2.2. Current Semantic Segmentation Network Approach

The general semantic segmentation network consists of an encoder and a decoder. Encoder is a pre-trained network, including AlexNet, VGG-Net, ResNet, etc. Decoder projects the discriminable characteristics into the pixel space to obtain the intensive classification.[5]. At present, some classical semantic segmentation network methods generally include Unet, AD-LinkNet and DeepLab.

Unet[6] is an optimized semantic segmentation network based on FCNs, which is composed of two parts. The first part is feature extraction, and the second part is up-sampling. However, the biggest difference between Unet and other semantic segmentation network is that Unet adopts the feature fusion method of "channel dimension splicing and fusion" to form thicker features.

AD-LinkNet[7] is a new semantic segmentation network based on the D-LinkNet and the integration of various network advantages. Its innovation lies in the addition of series-parallel combination, extended convolution and attention mechanism in the network, which enables it to complete more sophisticated semantic segmentation tasks. AD-LinkNet uses pre-trained ResNet as its encoder in part A, which can significantly increase the ability of representation and generalization. AD-LinkNet also introduces several supervised branches in order to establish a multi-task model in part B. So that different one-dimensional vectors lengths can be employed on different tasks. However, the decoder part of the AD-LinkNet is the same as LinkNet in part C.

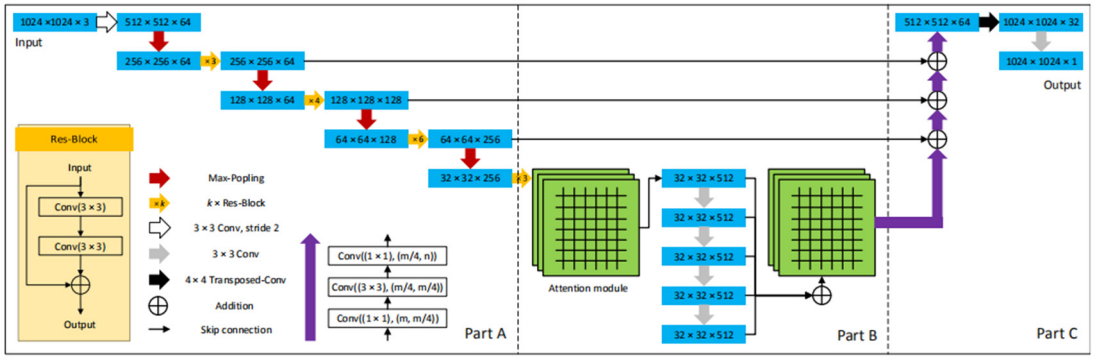

Figure 2. The structure diagram of AD-LinkNet[7]

The DeepLab V3 network[8] uses a spatial pyramid structure with stacked hole convolutions, also known as Atrous Convolution, which enables the convolutional layer to increase the corresponding receptive field of convolution without reducing the spatial dimension or increasing the number of parameters of this link, so as to enhance and improve the segmentation effect of the network. At the same time, V3 improves the ASPP module and references the idea of Hybrid Dilated Convolution(HDC)[9] which is used to mitigate the influence of "gidding issue" caused by the expanded convolution and expand the receptive field to aggregate global information, but the backbone is still ResNet101.

3. Backbone of Network

3.1. VGG-Net as backbone

The key to the great breakthrough of VGG lies in the selection of a smaller convolution kernel. Compared with AlexNet's 11x11 and LeNets's[10] 7x7 convolution kernel, VGG's 3x3 convolution kernel can not only consume less computation and introduce more non-linearity when obtaining the same receptive field, but also make the model more powerful in fitting ability. At the same time, small convolution kernel is more convenient to optimize convolution calculation [3]. Another innovation of VGG is the reference to Network in Network[11], which uses the convolution kernel of 1x1 to carry out non-linear processing through ReLU without affecting the dimension of input and output, so as to improve the nonlinearity of the model. Therefore, stronger fitting ability, simpler calculation and nonlinear optimization make it better as the backbone of the semantic segmentation network.

3.2. ResNet as backbone

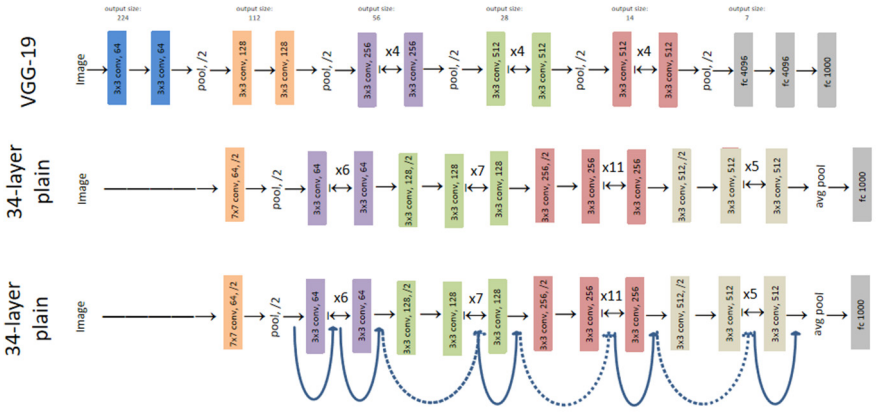

Figure 3. Example networks architecture for ImageNet

The figure compares VGG-19 and 34-layer ResNet [12]. There are two major differences: first, the kernel cannel of ResNet is much less than that of VGG-19, so even a large Kernel size can control the computation without reducing the receptive field. As a result, in the figure, when VGG-19 uses four 3x3 convolution cores, ResNet only needs to use one 7x7 convolution kernel. The second difference is that ResNet abandons FC and adopts Ave Pool instead, which can not only prevent overfitting, but also greatly improve the precision. Even in GoogleNet, which has not completely replaced FC in the previous generation, the precision of top-1 can be improved by almost 0.6%[13], which makes ResNet become one of the most excellent backbone-building networks at present.

3.3. Xception as backbone

The structure of Xception[14] is optimized on the basis of ResNet, but the convolutional layer in ResNet is replaced by Depthwise Separable Convolution[15]. In this module, the convergence process is faster and the accuracy is higher between the 3x3 convolution for learning spatial correlation and the 1x1 convolution for learning inter-channel correlation, when a non-linear activation function is not used. In the comparison of ImageNet classification performance, Xception also showed more outstanding performance than VGG-16 and ResNet-152 [14].

Table 1. Classification performance comparison on ImageNet (single crop, single model). VGG-16 and ResNet-152 numbers are only included as a reminder. The version of Inception V3 being benchmarked does not include the auxiliary tower.[14]

	Top-1 accuracy	Top-5 accuracy
VGG-16	0.715	0.901
ResNet-152	0.770	0.933
Inception V3	0.782	0.941
Xception	**0.790**	**0.945**

However, the potential problem is that although Depthwise Separable Convolution can improve the accuracy or reduce the theoretical computation, it is not very efficient in the existing convolutional neural network due to its fragmented computation process.

4. Evaluation

4.1. Introduction to data set

In this paper, CVPR DeepGlobe's road extraction data set[16] was selected, including urban, rural, rural, coastal, tropical rainforest and other scenes. Due to the small proportion of roads in the whole picture, roads are similar to rivers, railways and other shapes, and there is communication and interaction between roads and roads, and there may be a variety of obstacles blocking, which requires a high degree of accuracy in the semantic segmentation network.

4.2. Model effect comparison

Table 2. Model Test Result

Network Structure	IoU Score (%)
Unet-VGG-16	62.94
AD-LinkNet-ResNet101	63.37
AD-LinkNet-ResNet34	64.73
AD-LinkNet-Xception	64.81

IoU, as a very important function of the performance mAP calculation of the target detection algorithm, represents the coincidence degree between the predicted boundary and the actual boundary, which can directly reflect the performance of the algorithm. In this paper, VGG-16 is used as the backbone of Unet, and ResNet34, ResNet101 and Xception are used as the backbone of AD-LinkNet. Through experiments, their IoU scores are 62.94%, 64.73%, 63.82% and 64.81% respectively. After analyzing the above results, it can be found that: when VGG-16 is used as the backbone of the semantic segmentation network, it can extract long and wide roads with good effect, but its performance is not very good when faced with complex road extraction data set like the one we used in the experiment. As the backbone of the semantic segmentation network, ResNet34 achieves better results because it can extract better results from smaller roads. The road extraction data set used in the experiment, including a large number of narrow roads and bridges, just conforms to the characteristics of ResNet34. As a deeper network than ResNet34, ResNet101 should achieve better performance in theory, but the experimental results are just the opposite. This paper believes that it is because the data volume of the data set used in the experiment is too small. When Xception is used as the backbone of AD-LinkNet, it achieves the best effect, because Xception is optimized on the basis of ResNet, which makes Xception not only inherit ResNet's advantage in the face of small road extraction, but also show its ability to extract objects when covered by occlusions. Therefore, we believe that Xception as the backbone can provide the best performance for the semantic segmentation network in the face of complex interleaved and multiple occlusions.

5. Conclusion

The continuous extension and expansion of deep learning has laid a foundation for semantic segmentation network, and the continuous optimization of classification network has gradually improved the accuracy of semantic segmentation. This paper reviews a variety of classification networks like AlexNet and VGG. Then discusses the differences VGG, ResNet and Xception when they are used as backbone. We then use CVPR DeepGlobe's road extraction data set and compare the performance of multiple backbones of semantic segmentation network through experiments. Through experimental comparison, Xception showed more outstanding performance than ResNet34, ResNet101 and VGG. Ultimately, we hope to be able to try more excellent, advanced and newer backbones for semantic segmentation network.

6. References

[1] Krizhevsky, A. , Sutskever, I. , & Hinton, G. . (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS (Vol.25). Curran Associates Inc.
[2] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115 (3), 211-252.
[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. ArXiv preprint arXiv: 1409.1556, 2014. 2, 5
[4] He, K., Zhang, X., Ren, S., & Sun, J .. (2015). Deep residual learning for image recognition.
[5] Dinh Viet Sang, Nguyen Duc Minh. "Fully Residual Convolutional Neural Networks for Aerial Image Segmentation", Proceedings of the Ninth International Symposium on Information and Communication Technology - SoICT 2018, 2018
[6] Ronneberger, O., Fischer, P., & Brox, T .. (2015). U-net: convolutional networks for biomedical image segmentation.
[7] Wu, M., Zhang, C., Liu, J., Zhou, L., & Li, X. (2019). Towards Accurate High Resolution Satellite Image Semantic Segmentation. IEEE Access, 7, 55609-55619 .
[8] Chen Tianhua, Zheng Siqun, & Yu Junchuan. (2018). Remote sensing image segmentation using improved deeplab network. Measurement and control technology, 37 (11), 40-45.
[9] Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., & Hou, X., et al. (2017). Understanding convolution for semantic segmentation.

[10] Lecun, Y. , Bottou, L. , Bengio, Y. , & Haffner, P. . (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[11] Lin, Min, Qiang Chen, and Shuicheng Yan. Network In Network. arXiv:1312.4400 [Cs], December, 2013.
[12] He, K., Zhang, X., Ren, S., & Sun, J .. (2015). Deep residual learning for image recognition.
[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., & Anguelov, D., et al. (2014). Going deeper with convolutions.
[14] Chollet, François. (2016). Xception: deep learning with depthwise separable convolutions.
[15] Howard, A. G. , Zhu, M. , Chen, B. , Kalenichenko, D. , Wang, W. , & Weyand, T. , et al. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications.
[16] Demir, I. , Koperski, K. , Lindenbaum, D. , Pang, G. , Huang, J. , & Basu, S. , et al. (2018). Deepglobe 2018: a challenge to parse the earth through satellite images.