D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction

Lichen Zhou, Chuang Zhang, Ming Wu
Beijing University of Posts and Telecommunications
{zhoulichen, zhangchuang, wuming}@bupt.edu.cn

Abstract

Road extraction is a fundamental task in the field of remote sensing which has been a hot research topic in the past decade. In this paper, we propose a semantic segmentation neural network, named D-LinkNet, which adopts encoder-decoder structure, dilated convolution and pretrained encoder for road extraction task. The network is built with LinkNet architecture and has dilated convolution layers in its center part. Linknet architecture is efficient in computation and memory. Dilation convolution is a powerful tool that can enlarge the receptive field of feature points without reducing the resolution of the feature maps. In the CVPR DeepGlobe 2018 Road Extraction Challenge, our best IoU scores on the validation set and the test set are 0.6466 and 0.6342 respectively.

1. Introduction

Road extraction from satellite images has been a hot research topic in the past decade. It has a wide range of applications such as automated crisis response, road map updating, city planning, geographic information updating, car navigations, etc. In the field of satellite image road extraction, a variety of methods have been proposed in recent years. Most of these methods can be seperated into three categories: generating pixel-level labeling of roads [1, 2], detecting skeletons of roads [3, 4] and a combination of both [5, 6].

In the DeepGlobe Road Extraction Challenge [7], the task of road extraction from satellite images was formulated as a binary classification problem: to label each pixel as road or non-road. In this paper, we handling the road extraction task as a binary semantic segmentation task to generate pixel-level labeling of roads,.

Recently, deep convolutional neural networks (DCNN) [8, 9, 10, 11] have shown their dominance on many visual recognition tasks. In the field of image semantic segmentation, fully-convolutional network (FCN) [12] architecture, which can produce a segmentation map for an entire input image through single forward pass, is prevalent. Most latest excellent semantic segmentation networks [13, 14, 15, 16] are improved versions of FCN.

Several previous works have applied deep learning to road segmentation task. Mnih and Hinton [17] employed restricted Boltzmann machines to segment road from high resolution aerial images. Saito *et al*. [18] used a classification network to assign each patch extracted from the whole image as road, building or background. Zhang *et al*. [1] followed the FCN architecture and employed a Unet with residual connections to segment roads from one image through single forward pass. In this paper, we follow these methods, using DCNN to handle road segmentation task.

Although has been extensively studied in the past years, road segmentation from high resolution satellite images is still a challenging task due to some special features of the task. First, the input images are of high-resolution, so networks for this task should have large receptive field that can cover the whole image. Second, roads in satellite images are often slender, complex and cover a small part of the whole image. In this case, preserving the detailed spacial information is significant. Third, roads have natural connectivity and long span. Taking these natural properties of roads in consideration is necessary. Based on the challenges discussed above, we propose a semantic segmentation network, named D-LinkNet, which can properly handle these challenges.

D-LinkNet uses Linknet [15] with pretrained encoder as its backbone and has additional dilated convolution layers in the center part. Linknet is an efficient semantic segmentation neural network which takes the advantages of skip connections, residual blocks [10] and encoder-decoder architecture. The original Linknet uses ResNet18 as its encoder, which is a pretty light but outperforming network. Linknet has shown high precision on several benchmarks [19, 20], and it runs pretty fast.

Dilated convolution is a useful kernel to adjust receptive fields of feature points without decreasing the resolution of feature maps. It was widely used recently, and it

6회

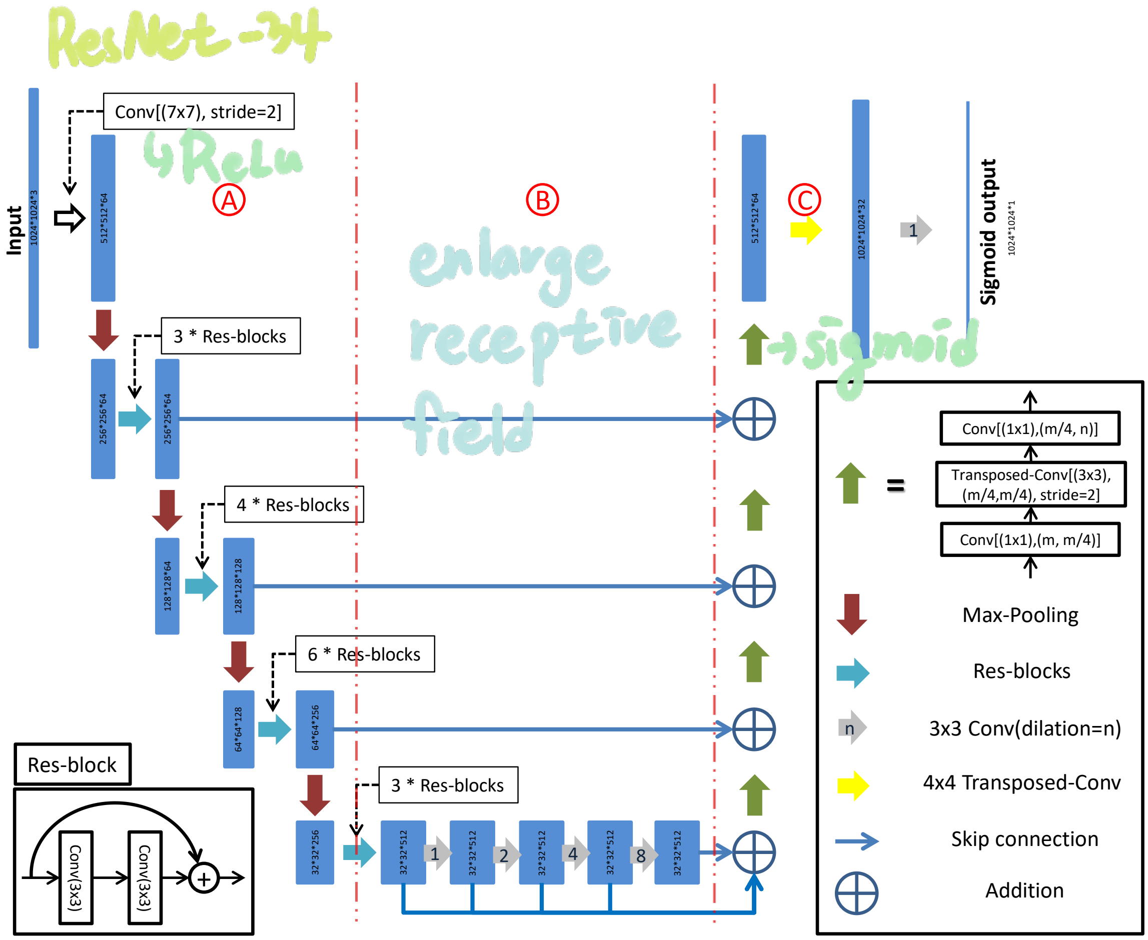

Figure 1. D-LinkNet architecture. Each blue rectangular block represents a multi-channel features map. Part A is the encoder of D-LinkNet. D-LinkNet uses ResNet34 as encoder. Part C is the decoder of D-LinkNet, it is set the same as LinkNet decoder. Original LinkNet only has Part A and Part C. D-LinkNet has an additional Part B which can enlarge the receptive field and as well as preserve the detailed spatial information. Each convolution layer is followed by a ReLU activation except the last convolution layer which use sigmoid activation.

every conv

last conv

generally has two types, cascade mode like [21] and parallel mode like [16], both modes have shown strong ability to increase the segmentation accuracy. We take advatages of both modes, using shortcut connection to combine these two modes.

Transfer learning is a useful method that can directly improve network preformance in most situation [22], especiall when the training data is limited. In semantic segmantation field, initializing encoders with ImageNet [23] pretrained weights has shown promising results [16, 24].

In the DeepGlobe Road Extraction Challenge, our best single model got IoU score of 0.6412 on the validation set.

2. Method

2.1. Network Architecture

In the DeepGlobe Road Extraction Challenge, the original size of the provided images and masks is 1024×1024, and the roads in most images span the whole image. Still, roads have some natural properties such as connectivity, complexity *et al*. Considering these properties, D-LinkNet is designed to receive 1024×1024 images as input and preserve detailed spacial information. As shown in Figure 1, D-LinkNet can be split in three parts A, B, C, named encoder, center part and decoder respectively.

D-LinkNet uses ResNet34 [10] pretrained on ImageNet [23] dataset as its encoder. ResNet34 is originally designed for classification task on mid-resolution images of size 256×256, but in this challenge, the task is to segment roads from high-resolution satellite images of size 1024×1024. Considering the narrowness, connectivity, complexity and long span of roads, it is important to increase the receptive field of feature points in the center part of the network as well as keep the detailed information. Using pooling layers could multiply increase the receptive field of feature points, but may reduce the resolution of center feature maps and drop spacial information. As shown by some state-of-the-art deep learning models [21, 25, 26, 16], deliated rate

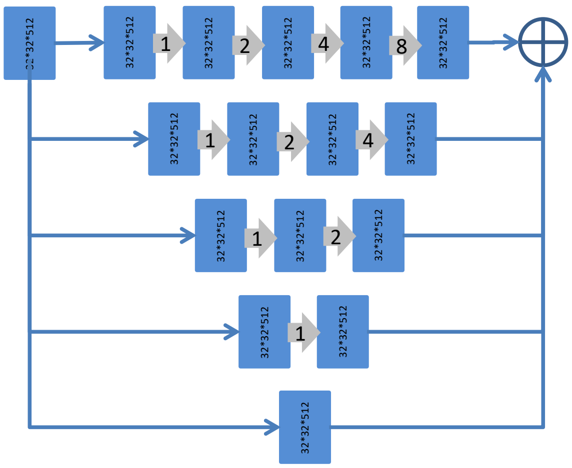

Figure 2. The center dilation part of D-LinkNet can be unrolled as this structure. It contains dilated convolution both in cascade mode and parallel mode, and the receptive field of each path is different, so the network can combine features from different scales. From top to bottom, the receptive fields are 31, 15, 7, 3, 1 respectively.

dilated convolution layer can be desirable alternative of pooling layer. D-LinkNet uses several dilated convolution layers with skip connections in the center part.

Dilated convolution can be stacked in cascade mode. As shown in the Figure1 of [21], if the dilation rates of the stacked dilated convolution layers are 1, 2, 4, 8, 16 respectively, then the receptive field of each layer will be 3, 7, 15, 31, 63. The encoder part (RseNet34) has 5 downsampling layers, if an image of size 1024×1024 go through the encoder part, the output feature map will be of size 32×32. In this case, D-LinkNet uses dilated convolution layers with dilation rate of 1, 2, 4, 8 in the center part, so the feature points on the last center layer will see 31×31 points on the first center feature map, covering main part of the first center feature map. Still, D-LinkNet takes the advantage of multi-resolution features, and the center part of D-LinkNet can be viewed as the parallel mode as shown in Figure 2.

The decoder of D-LinkNet remains the same as the original LinkNet [15], which is computationally efficient. The decoder part uses transposed convolution [27] layers to do upsampling, restoring the resolution of feature map from 32×32 to 1024×1024.

2.2. Pretrained Encoder

Transfer learning is an efficient method for computer vision, especially when the number of training images is limited. Using ImageNet [23] pretrained model to be the encoder of the network is a method widely used in semantic segmentation field [16, 24]. In the DeepGlobe Road Extraction Challenge, we found that transfer learning can accelerate our network convergence and make it have better performance with almost no extra cost.

3. Experiments

In the DeepGlobe Road Extraction Challenge. We use PyTorch [28] as the deep learning framework. All models are trained on 4 NVIDIA GTX1080 GPUs.

3.1. Dataset

We test our method on DeepGlobe Road Extraction dataset [7], which consists of 6226 training images, 1243 validation images and 1101 test images. The resolution of each image is 1024×1024. The dataset is formulated as a binary segmentation problem, in which roads are labeled as foreground and other objects are labeled as background.

3.2. Implementation details

In the training phase, we did not use cross validation[1]. Still, we wanted to make full use of the provided data, so we trained our model on all of the 6226 labeled images, and only used the 1243 validation images provided by the organizer for validation. This may be at the risk of overfiting on the training set, so we did data augmentation in an ambitious way, including horizontal flip, vertical flip, diagonal flip, ambitious color jittering, image shifting, scaling.

For our best model, we used BCE (binary cross entropy) + dice coefficient loss as loss function and chose Adam [29] as our optimizer. The learning rate was originally set 2e-4, and reduced by 5 for 3 times while observing the training loss decreasing slowly. The batch size during training phase was fixed as 4. It took about 160 epochs for our network to converge.

We did test time augmentation(TTA) in the predicting phase, including image horizontal flip, image vertical flip, image diagonal flip (predicting each image $2 \times 2 \times 2 = 8$ times), and then restored the outputs to the match the origin images. Then, we averaged the prob of each prediction, using 0.5 as our prediction threshold to generate binary outputs.

3.3. Results

During the DeepGlobe Road Extraction Challenge, we trained a deep Unet with 7 pooling layers, which can cover images of size 1024×1024, as our baseline model, and trained a LinkNet34 with pretrained encoder but without dilated convolution in the center part. The performances of different model are shown in Table 1. We found that the pretrained LinkNet34 was just a little bit better than the Unet trained from scratch. We evaluated the IoU of masks predicted by Unet and masks predicted by LinkNet34, and

[1] It took about 40 hours for us to train one model, if we train models with 5-fold cross validation, it will take us 200 hours to try one architecture (too long for us), so we just dropped cross validation.

Model	IoU on validation set
Unet (7 pooling layers, no-pretrain)	0.6294
LinkNet34 (pretrained encoder)	0.6300
Ensemble of Unet and LinkNet34	0.6394
D-LinkNet (pretrained encoder)	0.6412

Table 1. Results on validation set of different models in the DeepGlobe Road Extraction Challenge. LinkNet34 with pretrained encoder got almost the same score as Unet on the validation set. D-LinkNet get higher score than the Ensembling of Unet and LinkNet34 on the validation set.

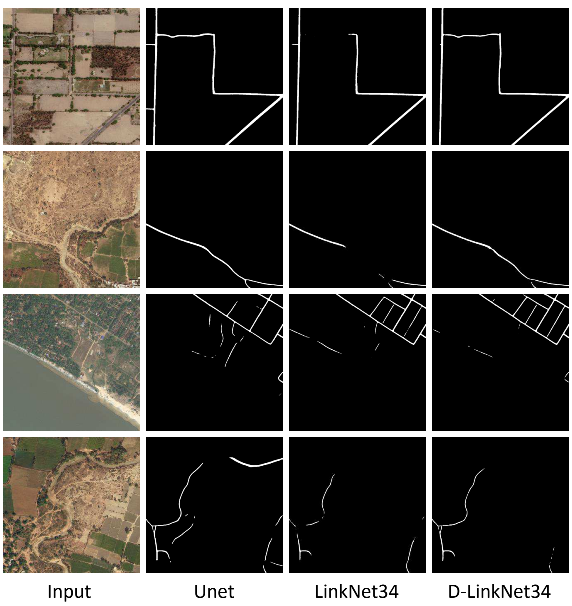

Figure 3. Example results of three models. The first two lines are examples showing the road connectivity problem in LinkNet34. There are several road interruptions in LinkNet34 results. The last two lines are examples showing the incorrection predicting of Unet. Unet is more likely to wrongly recognize roads as background or recognize something non-road like rivers as roads. D-LinkNet avoids weaknesses in Unet and LinkNet34, and makes better predictions.

found that on the validation set, the averaged IoU of these two models was 0.785, which we considered as a pretty low score. We thought these two models might get almost the same score in different ways. Our baseline Unet had larger receptive field but had no pretrained encoder and the center feature map's resolution was 8×8, which is too small to preserve detailed spacial information. LinkNet34 had pretrained encoder which made the network has better representation, but it only had 5 downsampling layers, hardly covering the 1024×1024 images. While reviewing the outputs from these two models, we found that although LinkNet34 was better than Unet while judging an object to be road or not, it had road connectivity problem. Some examples are shown in Figure 3. By adding dilated convolution with shortcuts in the center part, D-LinkNet can obtain larger receptive field than LinkNet as well as preserve detailed information at the same time, and thus alleviated the road connectivity problem occurred in LinkNet34.

3.4. Analysis

We used several methods during the DeepGlobe Road Extraction Challenge, and we have done several experiments to find the contribution of each method. The most contributing method is test time augmentation(TTA), it contributes about 0.029 points. Using BCE + dice coefficient loss is better than BCE + IoU loss about 0.005 points. Pretrained encoder contributes about 0.01 points. Dilated convolution in the center part contributes about 0.011 points. Ambitious data augmentation is better than normal data augmentation without color jittering and shape transfromation about 0.01 points.

4. Conclusion

In this paper, we have proposed a semantic segmentation network, named D-LinkNet, for high resolution satellite imagery road extraction. By enlarging the receptive field and ensembling multi-scale features in the center part while keeping the detailed information at the same time, D-LinkNet can handle roads' properties such as narrowness, connectivity, complexity and long span to some extent. However, D-LinkNet still has the wrong recognition and road connectivity problems, we plan to do more research on these problems in the feature.

In addition, although the proposed D-LinkNet architecture was originally designed for the road segmentation task, we anticipate it may also be useful in other segmentation tasks, and we plan to investigate this in our future research.

References

[1] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. In *IEEE Geoscience and Remote Sensing Letters*. IEEE, 2018. 1

[2] Rasha Alshehhi and Prashanth Reddy Marpu. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. In *ISPRS journal of photogrammetry and remote sensing*, volume 126, pages 245–260. Elsevier, 2017. 1

[3] Bo Liu, Huayi Wu, Yandong Wang, and Wenming Liu. Main road extraction from zy-3 grayscale imagery based on directional mathematical morphology and vgi prior knowledge in urban areas. In *PloS one*, volume 10, page e0138071. Public Library of Science, 2015. 1

[4] Chinnathevar Sujatha and Dharmar Selvathi. Connected component-based technique for automatic extraction of road centerline in high resolution satellite images. In *EURASIP*

Journal on Image and Video Processing, volume 2015, page 8. Springer, 2015. 1

[5] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. *arXiv preprint arXiv:1802.03680*, 2018. 1

[6] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *International Conference on Computer Vision*, volume 2, 2017. 1

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018. 1, 3

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

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

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[11] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017. 1

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[14] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In *IEEE transactions on pattern analysis and machine intelligence*, volume 39, pages 2481–2495. IEEE, 2017. 1

[15] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. *arXiv preprint arXiv:1707.03718*, 2017. 1, 3

[16] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 1, 2, 3

[17] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pages 210–223. Springer, 2010. 1

[18] Shunta Saito, Takayoshi Yamashita, and Yoshimitsu Aoki. Multiple object extraction from aerial imagery with convolutional neural networks. In *Electronic Imaging*, volume 2016, pages 1–9. Society for Imaging Science and Technology, 2016. 1

[19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[20] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer, 2008. 1

[21] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2, 3

[22] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1717–1724. IEEE, 2014. 2

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2, 3

[24] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018. 2, 3

[25] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 2

[26] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition*, volume 1, 2017. 2

[27] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE, 2011. 3

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3