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

Yufeng Jiang

Abstract

In this paper, authors 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 conputatuion 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.

1. Introduction

Recently, deep convolutional neural networks (DCNN) [6, 9] have shown their dominance on many visual recognition tasks. In the field of image semantic segmentation, fully-convolutional network (FCN) [7] architecture, which can produce a segmentation map for an entir input image through single forward pass, is prevalent. Most latest excellent semantic segmentation networks [8] are improved versions of FCN. In this paper, they use 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. Based on these challenges, they propose a semantic segmentation network, named D-LinkNet, which can properly handle these challenges.

D-LinkNet uses Linknet [1] with pretrained encoder as its backbone and has additional dilated convolutional layers in the center part. Linknet is an efficient semantic segmentation neural network which takes the advantages of skip connections, residual bolcks and encoder-decoder architecture. The original Linknet uses ResNet18 as its encoder, which is a pretty light but outperforming network. Linknet has shown hhigh precision on several benchmarks, 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 generally has two types, cascade mode like [10] and parallel model like [2], both modes have shown strong ability to increase the segmentation accuracy. They takd advatages of both modes, using shortcut connection to combine these two modes.

2. Method

2.1. Network Architecture

Because the original size of the provided images and masks is 1024×1024, D-LinkNet is designed to receive 1024×1024 iamges as input and preserve detailed spacial information. As shwon 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 [5] pretrained on ImageNet [4] dataset as its encoder. ResNet34 is originally designed for classification task on mid-resolution images of size 256×256, but in the challenge they are faced, 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 netwrok 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 shwon by some state-of-the-art deep learning models, dilated convolution layer can be desirable alternative of pooling layer. D-LinkNet uses several dilated convolution layers with skip connections in the center part.

2.2. Pretrained Encoder

Transfer learning is an efficient method for computer vision, especially when the number of training images is limited. Using ImageNet [4] pretrained model to be the encoder of the network is a method widely used in semantic segmentation field. The transfer learning can accelerate their network convergnece and make it have better performance with almost no extra cost.

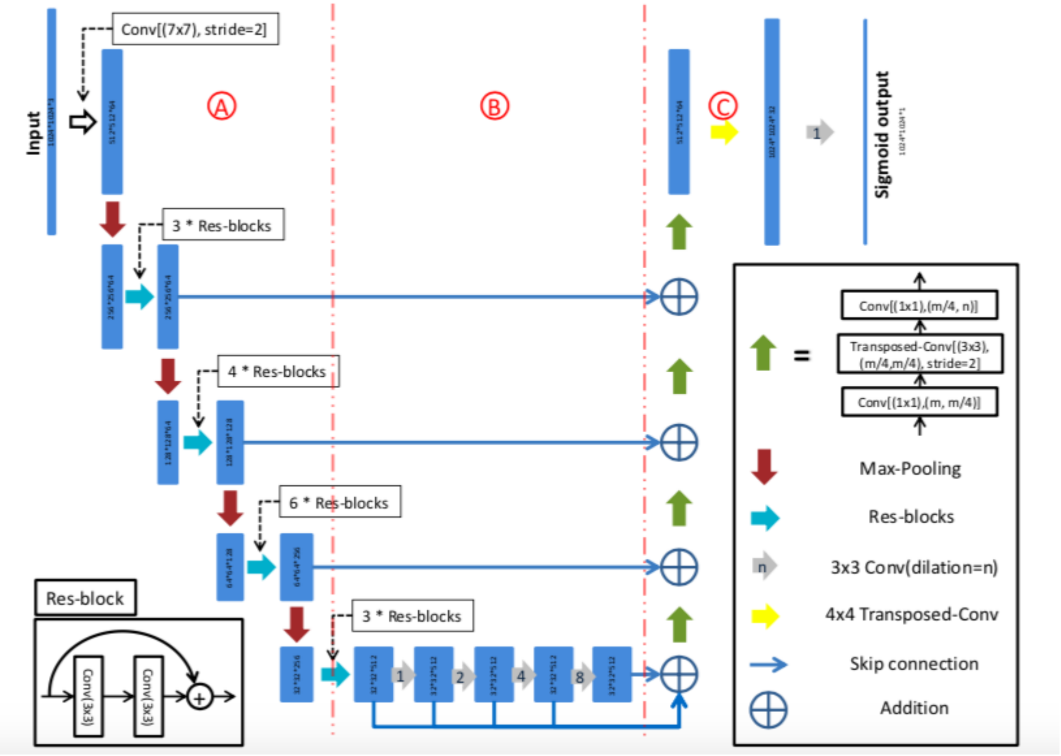

Figure 1. D-LinkNet architecture. Each blue rectangular block represents a multi-channel features map. Part A is the encoder of D-LinkNet. D-LinkNet uses ResNet34 as encoder. Part C is the decoder of D-LinkNet, it is set the same as LinkNet decoder. Original LinkNet only has Part A and Part C. D-LinkNet 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.

3. Experiments

They use PyTorch as the deep learning framework. All models are trained on 4 NVIDIA GTX1080 GPUs.

3.1. Dataset

They test their method on DeepGlobe Road EXtraction dataset [3], which consists of 6226 training images, 1243 validation images and 1101 test images. The resolution of each imges 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, they did not use cross validation[1]. Still, they wanted to make full use of the provided data, so they trained their model on all of the 6226 labeled iamges, and only used the 1243 validation images provided by the organizer for validation.

Model	IoU on validation set
Unet (7 pooling layers, no-pretrain)	0.6294
LinkNet34 (pretrained encoder)	0.6300
Ensembele 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.

3.3. Results

They trained a deep Unet with 7 pooling layers, which can cover images of size 1024 × 1024, as their 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.

They found that the pretrained LinkNet34 was just a little bit better than the Unet trained from scratch. They evaluated the IoU of masks predicted by Unet and masks pre-

[1] It took about 40 hours for us to train one model, if they 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.

dicted by LinkNet34, and found that on the validation set, the averaged IoU of thesee two models was 0.785, which they considered as a pretty low score.

References

[1] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. *arXiv preprint arXiv:1707.03718*, 2017. 1

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 1

[3] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, and K. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[7] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *CVPR*, 2015. 1

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

[10] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1