Accurate road segmentation in remote sensing images using dense residual learning and improved focal loss

Junjun Ma¹, Zhaohong Xu², Ergong Zheng²* and Qiongjian Fan²

¹ Institute of Information and Navigation, Air Force Engineering University, Xi’an, Shanxi, 710077, China
² Aviation University of Air Force, Changchun, Jilin, 130022, China
*Corresponding author’s e-mail: 330983114@qq.com

Abstract. In this paper, we propose a road segmentation model using deep learning technology. The model is essentially a U-net with pre-trained DenseNet-169 encoder. Multi-scale and high-level semantic information are extracted effectively by dense residual learning and attention mechanism. In addition, an improved focal loss function is proposed to handle extremely imbalanced road samples. Experimental results demonstrate that our proposed road segmentation model can accurately extract complex road areas in remote sensing images and has IoU of 0.6308 which ranks 22st in the DeepGlobe Road Extraction Challenge leaderboard at the time of writing.

1. Introduction

Distribution of road network is the key element of digital map in the information age. Accurate road extraction plays a very important role in urban development planning, geographic information management, positioning and navigation, and disaster prevention. Road extraction can be regarded as an image segmentation problem. At present, common methods for road segmentation can be divided into three categories including methods based on encoding and decoding structure, information fusion and recurrent neural network. Representative models of these methods mainly include FCN[1], SegNet[2], U-net[3], PSPNet[4], FPN[5], DeepLab[6], ReSeg[7], etc. In recent years, a variety of methods have been proposed for the road segmentation task in remote sensing images. In [8], a semantic segmentation model is proposed by combining U-net model and residuals learning. In order to simplify the training process of model and reduce the parameters, mean square error loss function and residual structure are taken into training the model and extracting roads with end-to-end training. The work [9] uses the pretrained ResNet-34 network as the encoder of U-net, and uses the weighted summation of binary cross entropy loss and Jaccard loss to be the total loss function and significantly improved the accuracy of road segmentation based on U-net. Adapted from LinkNet [10], D-LinkNet [11] adds a multi-scale feature extraction module between encoder and decoder and obtains accurate road segmentation results through an effective method for extracting multi-scale and high-level semantic information.

There are three main characters for accurate road segmentation in aerial remote sensing images: 1) Remote sensing images have higher complexity than natural images, 2) Road has the characteristics of slender linear and large span, 3) Training samples are extremely imbalanced. To overcome above three difficulties, we proposed an improved deep encoder-decoder architecture. In this paper, our main contributions are:
Designing a new encoder-decoder model for accurate road segmentation in remote sensing images. Our proposed model based on DenseNet-169 backbone can automatically extract the hierarchical features from remote sensing images and fuse multi-level features to make dense predictions, and it can be trained end-to-end;

- Using a DenseASPP module with different dilated rates to extract multi-scale features of road areas. Dilated convolution can compute the responses of any layer with any desirable resolution without increasing computation complexity;
- Integrating attention mechanism and fusion strategy to reuse multi-layer features for better dense predictions;
- Proposing an improved and universal focal loss to deal with extremely imbalanced samples. As far as we know, this kind of improvement based on focal loss is the first effective attempt for road segmentation in remote sensing images.

2. Proposed model

2.1. Road segmentation framework

Figure 1. Road segmentation model based on DenseNet-169 network. The number in a dotted box or outside a dotted box denotes different channels.

Structure of our improved road segmentation model is shown in figure 1. We choose DenseNet-169[12] to be the encoder of U-net. DenseNet-169 is a deep convolution network based on dense residual connection structure, and it achieves better classification accuracy on ImageNet dataset with fewer parameters. In the decoder, we choose the same module used in LinkNet for up-sample, which is an efficient module with fewer parameters. To take advantage of low-level features, we add 1×1 convolution followed by the SE-Net module[13] in shortcut connection. SE-Net is essentially a kind of...
attention mechanism, which can adaptively re-calibrate channel-wise features by using global information to selectively emphasize informative features and suppress useless ones, boosting feature discriminability. Besides, instead of using only the last layer’s output, we add all the outputs of different decoder stages and use 1×1 convolution to get segmentation masks. Outputs of different decoder stages are processed by 1×1 convolution to obtain same channels before up-sample.

2.2. Dense residual learning module

Receptive field is very important to obtain contextual information. Dilated convolution [14] is an effective convolution method to increase the model’s receptive field without increasing the computation. In work [15], a DenseASPP module is proposed for extracting multi-scale and high-level semantic information efficiently. Different from the work [15], we choose cascaded convolution layers with dilated rate \( 2^{n-1} \) (\( n \) is the layer’s index) to avoid the grid effect[16]. All dilated convolution layers’ kernels are 3×3. The receptive field of cascaded dilated convolution layers can be calculated by equation (1).

\[
R = \sum_{i=1}^{n} (d_i - d_{i-1} + 1) - n + 1
\]

where \( R \) denotes the receptive field, \( d \) denotes the dilated rate, \( k \) denotes the kernel size, \( n \) denotes numbers of layers and \( i \) is index. It should be noted that cascaded number of dilated convolution layers is determined by its input size. To strengthen nonlinear mapping ability and reduce parameters, a bottleneck with kernel size 1×1 is added in front of every dilated convolution layer.

2.3. Improved loss function

Focal loss [17] is often used to solve the problem of imbalanced samples, so we also choose it in this paper. The standard binary cross entropy loss and focal loss are shown in equation (2) and equation (3) respectively.

\[
\text{Loss}_{\text{BCE}} = -p_t \log(p_t) \quad (t = 0,1)
\]

\[
\text{Loss}_{\text{FL}} = -\alpha (1-p_t)^{\gamma} \log(p_t) \quad (t = 0,1)
\]

where \( \alpha \) and \( \gamma \) are hyper parameters and \( p_t \) is the predicted probability.

We analyze the original focal loss function from two aspects. Firstly, samples which are very easy to classify contribute little to the optimization of segmentation model. Secondly, original focal loss does not use an effective method to increase loss contribution of samples which are relatively difficult to classify. Based on above two considerations, we believe that loss contribution of samples which are relatively easy to classify should be diminished and loss contribution of samples which are relatively difficult to classify should be increased simultaneously. Thus, we perform some adjustment for the original focal loss from equation (3) to equation (4).

\[
\text{Loss}_{\text{FL}}^{*} = \alpha (1-p_t)^{\gamma} \left( \log(p_t) \right)^2, \quad (t = 0,1)
\]

Standard binary cross entropy loss, original focal loss and our improved focal loss are shown in figure 2. As we can see in figure 2, \(-\log(p_t)\) can increase the total loss (the yellow area in figure 2(b)) when \( p_t \) is near to 0, which promote the model to focus more on those samples that are very difficult to classify. When \( p_t \) is near to 1, \(-\log(p_t)\) is so close to 0 that it can reduce the loss (the yellow area in figure 2(c)) contribution from samples which are very easy to classify. For road extraction task, we perform normalization processing for our improved focal loss based on pixel mean. Finally, our improved focal loss can be represented in equation (5).

\[
\text{Loss}_{\text{FL}}^{*} = \frac{\sum_{c=0}^{N_o} \left\{ \alpha (1-p_{t,c})^{\gamma} \left[ \log(p_{t,c}) \right]^2 + (1-\alpha) p_{t,c}^{\gamma} \left[ \log(1-p_{0,c}) \right]^2 \right\}}{N_o}
\]
where $p_{0,i}$ and $p_{1,i}$ are the $p_t$ when $t$ is 0 or 1, $\alpha$ and $\gamma$ are the default parameters in original focal loss and are 0.25 and 2.0 respectively.

In this paper, the weighted sum of improved focal loss and Jaccard loss is used for optimizing the network weights. For semantic segmentation, Jaccard loss can be shown in equation (6).

$$Loss_J = \log \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{y_i y_i^*}{(y_i + y_i^*) - y_i y_i^*} \right]$$  \hspace{1cm} (6)$$

where $y_i$ denotes the true value of the pixel, $y_i^*$ is the predicted value of the network, $n$ is the total number of pixels.

Total loss function is shown in equation (7).

$$Loss = \beta \cdot Loss_{FL} + (1 - \beta) \cdot Loss_{JL}$$  \hspace{1cm} (7)$$

where $\beta$ denotes the weight factor. Same as most people did [18], we set $\beta$ equal to 0.7.

3. Results

3.1. Experimental details

The experimental dataset is DeepGlobe Road Extraction Dataset[19]. The numbers for training, validation and test in this dataset are 6226, 1243 and 1101 respectively, and there is no mask in validation set of this dataset. All the training images are RGB images with ground resolution of 0.5 meters and corresponding road segmentation masks.

In the training phase, we have done lots of data augmentation processing which include horizontal and vertical flip, random image rotation, image shifting and scaling, ambitious color jittering and HSV color-space transformation to avoid the risk of overfitting. All the experiments are implemented based on Keras. We choose the Adam optimizer with initial learning rate 0.0001 and the drop-based learning
rate schedule. We have done test time augmentation (TTA) in the predicting phase, including image horizontal, vertical and diagonal flip and image rotation with 90, 180, 270 degree, and then restored the outputs to the match the origin images. Then, we averaged these outputs after being restored to generate final segmentation. At last, we choose IoU criterion[20] to evaluate our segmentation model.

3.2. Loss function experiments
IoU criterion of proposed model trained with standard binary cross entropy loss, original focal loss and our improved focal loss are shown in table 1. We can see that our improved focal loss function gets the best IoU value compared with the other two loss functions.

3.3. Ablation experiments
We have performed ablation experiments on our proposed model. It should be noted that all the experiments are trained with our improved focal loss function. From table 2, we can see that MSP and DenseASPP play an important role in extracting multi-scale and high-level semantic information. It is obvious that our modules in table 2 are of significance, all of which increase the IoU criterion.

| Table 1. Loss function experimental results. |
|---------------------------------------------|
| Loss            | IoU(%) |
| Binary cross entropy | 58.56  |
| Focal loss       | 61.17  |
| Improved focal loss | 63.08  |

| Table 2. Ablation results. SN denotes shortcut connection with 1×1 convolution. MSP denotes model’s multi-scale dense prediction. |
|------------------------------------------------------------------------------------------------------------------|
| SN  | SE-Net | MSP          | DenseASPP | IoU(%) |
|-----|--------|--------------|-----------|--------|
| ✔   | ✔      | ✔            |           | 60.86  |
| ✔   | ✔      | ✔            |           | 61.02  |
| ✔   | ✔      | ✔            |           | 61.44  |
| ✔   | ✔      | ✔            | ✔         | 62.66  |
| ✔   | ✔      | ✔            | ✔         | 63.08  |

| Table 3. Comparison experiments of different models. |
|-----------------------------------------------------|
| Models   | Backbone     | Params | IoU(%) |
|----------|--------------|--------|--------|
| U-net    | DensNet-169  | 31M    | 54.05  |
| PSPNet   | DensNet-169  | 10M    | 58.78  |
| FPN      | DensNet-169  | 17M    | 61.62  |
| LinkNet  | DensNet-169  | 16M    | 61.98  |
| D-LinkNet34 | ResNet-34    | 119M   | 62.83  |
| D-LinkNet50 | ResNet-50    | 831M   | 63.42  |
| D-LinkNet101 | ResNet-101   | 904M   | 63.37  |
| U-net34  | ResNet-34    | -      | 64.00  |
| Proposed | DensNet-169  | 36M    | 63.08  |

3.4. Comparison experiments
Experiments for the standard U-net[3], PSPNet[4], FPN[5], LinkNet[10], D-LinkNet[11], U-net34 with ResNet-34 backbone[18] and our proposed model are shown in table 3. It is noted that only the standard U-net is trained with group normalization. As we can see in table 3, IoU of LinkNet is better than FPN and PSPNet. U-net34 has the highest IoU but has more parameters than our proposed model because of the backbone. The IoU of D-LinkNet with ResNet-50 backbone is higher than our model.
but its model size is extremely big. Parameters of our proposed model is similar to the standard U-net, but the IoU of our proposed model is about 9% higher than the standard U-net whose IoU is only 54.05% while ours is 63.08% which ranks 22st in the DeepGlobe Road Extraction Challenge leaderboard at the time of writing. Compared to the standard U-net, IoU of our proposed model has increased about 17%.

The road segmentation results can be seen in figure 3. Compared with other methods, our proposed model has two advantages: 1) edge of segmentation mask is more accurate, 2) roads that are shield by trees or buildings can still be segmented more accurately than other methods. The experimental results show that our proposed model has higher segmentation accuracy than other methods in remote sensing images.

4. Conclusion
In this paper, we have proposed a new semantic segmentation network for road extraction in remote sensing images. This approach mainly relies on three key factors. First, dense residual learning module with different dilated rates can extract the multi-scale and high-level semantic information. Second, attention mechanism and fusion strategy can reuse multi-layer features. Third, an improved focal loss function aims to optimize model with extremely imbalanced samples. At last, the best public score of our single model on the public leaderboard is 0.6308 which ranks 22st in the DeepGlobe Road Extraction Challenge leaderboard. In addition, we think that the edge loss of roads can help the model to find more accurate segmentation edge, which deserves further study.

Acknowledgments
This study was supported by the science and technology projects in Jilin Province Department of Education under grant No.2016-515.
References

[1] Long, J., Shelhamer, E., Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston. pp. 3431-3440.

[2] Badrinarayanan V., Kendall A., Cipolla R. (2017) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transaction on Pattern Analysis and Machine Intelligence, 39: 2481–2495.

[3] Ronneberger O., Fischer P., Brox T. (2015) U-Net: Convolutional networks for biomedical image segmentation. In: 18th International Conference on Medical Image Computing and Computer Assisted Interventions. Munich. pp. 234-241.

[4] Zhao H., Shi J., Qi X., Wang X. (2017) Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition. Hawaii. pp. 6230-6239.

[5] Seferbekov S., Iglovikov V., Buslaev A., Shvets A. (2018) Feature pyramid network for multi-class land segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 272-275.

[6] Chen L., Papandreou G., Kokkinos I., Murphy K., Yuille A. L. (2018) DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transaction on Pattern Analysis and Machine Intelligence, 40: 834–848.

[7] Visin F., Ciccone M., Romero A., et al. (2016) ReSeg: A recurrent neural network-based model for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas. pp. 426-433.

[8] Zhang Z., Liu Q., Wang Y., et al. (2018) Road extraction by deep residual U-Net. IEEE Geoscience and Remote Sensing Letters, 15:749–753.

[9] Yuan J. (2018) Learning building extraction in aerial scenes with convolutional networks. IEEE Transaction on Pattern Analysis and Machine Intelligence, 40: 2793–2798.

[10] Chaurasia A., Culurciello E. (2017) LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In: IEEE Visual Communications and Image Processing. St. Petersburg. pp. 518-521.

[11] Zhou L., Zhang C., Wu M. (2018) D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 192-196.

[12] Huang G., Liu Z., Maaten L. V. D., Weinberger K. Q. (2017) Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Hawaii. pp. 4700-4708.

[13] Hu J., Shen L., Sun G., et al. (2018) Squeeze-and-Excitation Networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 7132-7141.

[14] Wei X., Fu K., Gao X., et al. (2018) Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model. Remote Sensing Letters, 9: 199–208.

[15] Yang M., Yu K., Zhang C., et al. (2018) DenseASPP for semantic segmentation in street scenes. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 3684-3692.

[16] Wang Z., Ji S. (2018) Smoothed dilated convolutions for improved dense prediction. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. London. pp. 2486-2495.

[17] Lin T., Goyal P., Girshick R., et al. (2017) Focal loss for dense object detection. In: IEEE Conference on Computer Vision. Venice. pp. 2999-3007.

[18] Buslaev A., Seferbekov S., Iglovikov V., Shvets A. (2018) Fully convolutional network for automatic road extraction from satellite imagery. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 197-200.
[19] Demir I., Koperski K., Lindenbaum D., et al. (2018) DeepGlobe 2018: A challenge to parse the earth through satellite images. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 172-181.

[20] Pont-Tuset J., Marques F. (2016) Supervised evaluation of image segmentation and object proposal techniques. IEEE Trans. Pattern Analysis and Machine Intelligence, 38: 1465–1478.