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

DeepGlobe Workshop at CVPR 2018

Speaker: Ming Wu

Lichen Zhou, Chuang Zhang, Ming Wu, Ruihua Zhang
Beijing University of Posts and Telecommunications

Outline

- Introduction
- Solution Development
  - Unet
  - Test Time Augmentation
  - Ambitious Data Augmentation
  - LinkNet
- Final Approach
  - D-LinkNet
- Future Works

Task Description

- Automatically extracting roads and streets networks from satellite images.
- We formulated as a binary segmentation problem to detect all the road pixels in each area.

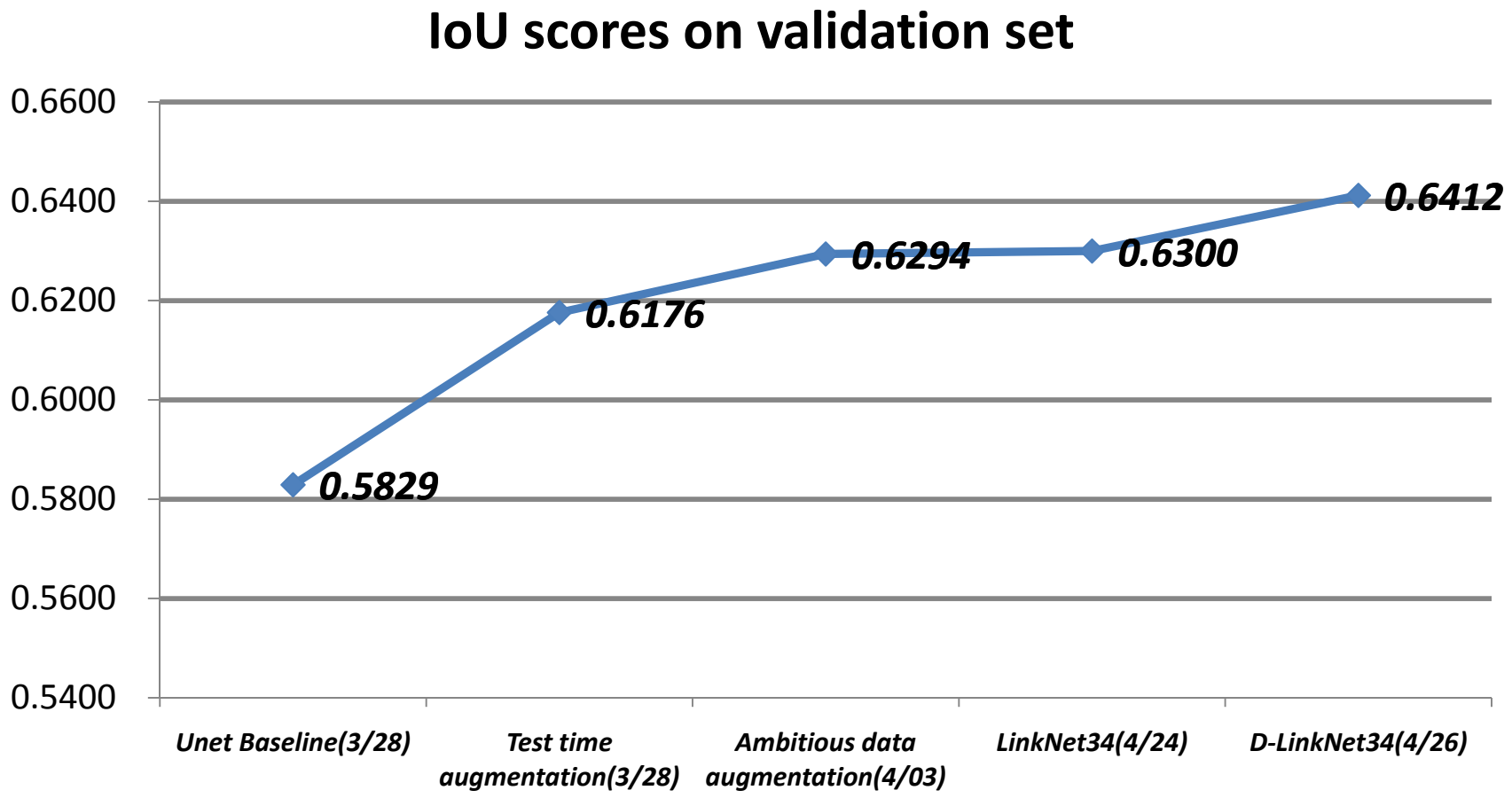
This is a Challenging Task

Features of the roads in satellite images :

- Slenderness
- Complexity
- Long span
- Connectivity

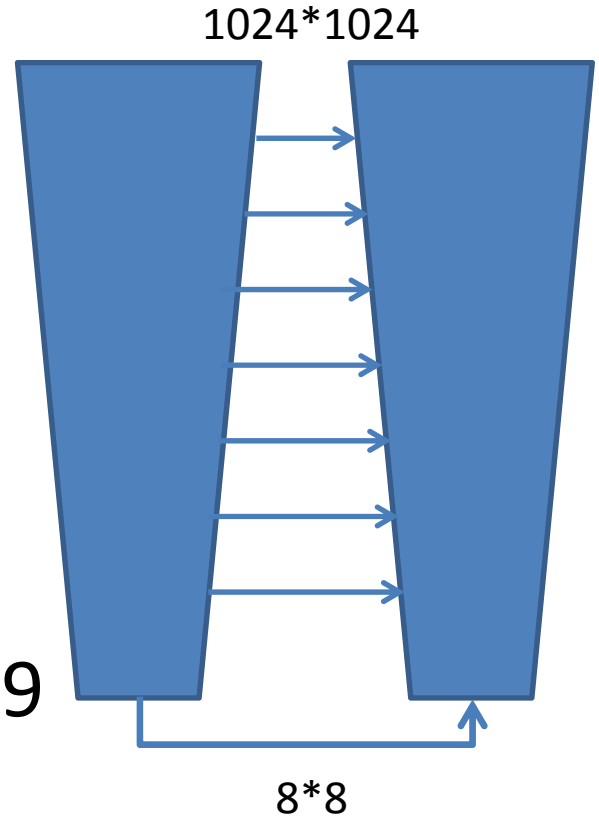
Illustrated on DEEPGLOBE – CVPR18 main page.

Solution Development



Unet Baseline (March 28)

- Unet is popular for segmentation of medical images and satellite images.
- Unet with 7 pooling layers
  - Input original resolution 1024*1024
  - Central feature map size is 8*8
  - Data aug: ver-flip, hor-flip, diag-flip
  - No cross validation
  - Loss_fn: BCE + Dice_coeff
- Scores on validation set: 0.5829



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

Loss Function and Optimizer

- BCE (Binary Cross Entropy) + Dice coefficient loss
- Adam[1] as our optimizer. Why not RMSProp[2]?
  - Adam converged faster than RMSProp on our pretrained model.
  - Our model trained by RMSProp is overfitting according to the experimental results

$$L = 1 - \frac{\sum_{i=1}^{N} |P_i \cap GT_i|}{\sum_{i=1}^{N} (|P_i| + |GT_i|)} + \sum_{i=1}^{N} BCELoss(P_i, GT_i)$$

Dice coefficient loss BCE(Binary Cross Entropy)

P is predicted image.
GT is annotation image.
N is batch size.

$$BCELoss(P, GT) = -\sum_{i=1}^{W}\sum_{j=1}^{H}[gt_{ij} \cdot \log p_{ij} + (1 - gt_{ij}) \cdot \log(1 - p_{ij})]$$

W is image width.
H is image height.
gt is a pixel in GT.
p is a pixel in P.

[1]Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[2] https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

Unet Baseline + TTA (March 28)

- Test Time Augmentation
  - ver-flip, hor-flip, diag-flip
- IoU scores on Validation Set

  0.5829 → 0.6176

Unet Baseline + TTA + Data Augmentation (April 3)

- Color Transfer(HSV)
  - ΔH ~ U(-15, 15)
  - ΔS ~ U(-15, 15)
  - ΔV ~ U(-30, 30)
- Spatial Transfer
  - Scale Ratio ~ U(-0.1, 0.1)
  - Aspect Ratio ~ U(-0.1, 0.1)
  - Shift Ratio ~ U(-0.1, 0.1)
- IoU scores on Validation Set

  0.6176 → 0.6294

Overfitting in Our Model

- Train Set : 6226 images
- Validation Set : 1243 images
- Test Set : 1101 images
- IoU score of our Unet model on training set is about 0.74

Road Extraction $\subset$ Pixel-wise Image Segmentation

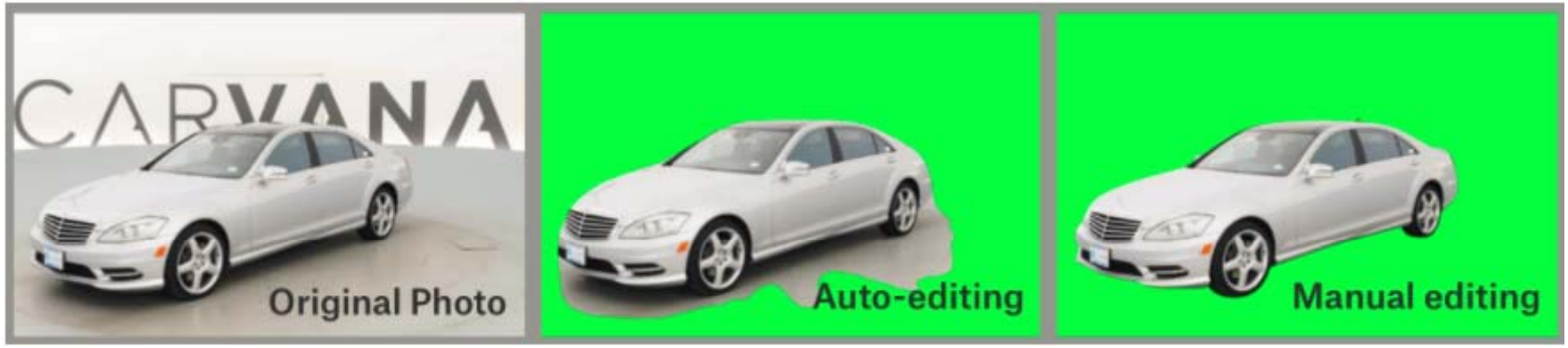
Kaggle Carvana Image Masking Challenge — Pixel-wise Image Segmentation

The Solutions of Winners

Alexander Buslaev:
- LinkNet+ResNet34
- ImageNet Pretrain

Artem Sanakoyeu:
- Unet from VGG11

Vladimi Iglovikov:
- Unet from VGG11
- ImageNet Pretrain



https://www.kaggle.com/c/carvana-image-masking-challenge

SpaceNet Challenge Road Detector —Pixel-wise Image Segmentation

The Solutions of Winners

Albu-solution:

- ResNet34 as encoder
- A Unet-like decoder

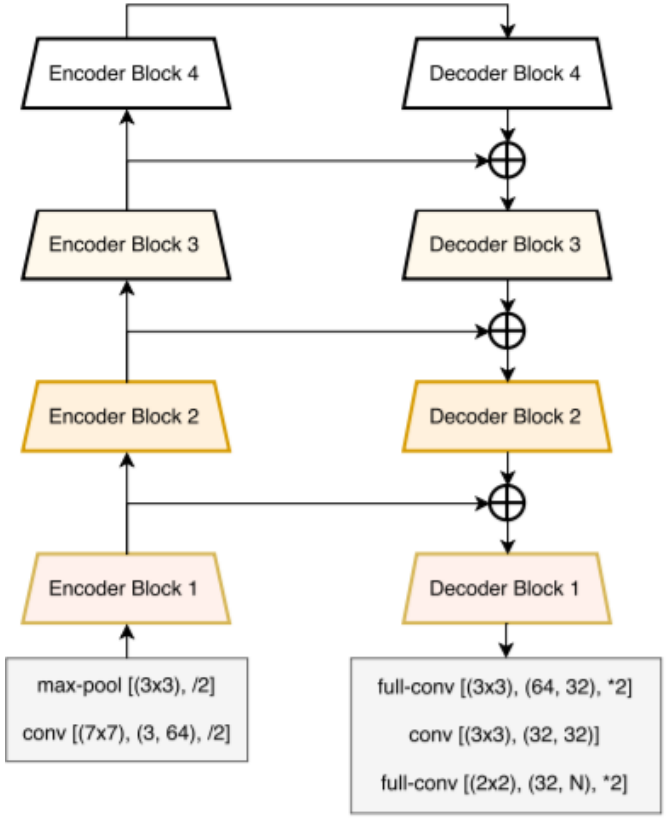
Cannab-solution:

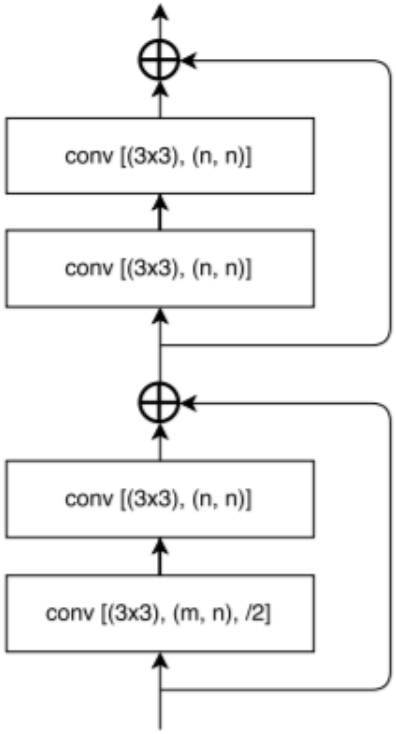
- Ensemble 12 Neural Network models
- Unet-like Neural Network with pretrained VGG16 as encoder
- LinkNet-like Neural Network with pretrained VGG16 as encoder
- ……

https://github.com/SpaceNetChallenge/RoadDetector
https://devblogs.nvidia.com/solving-spacenet-road-detection-challenge-deep-learning/

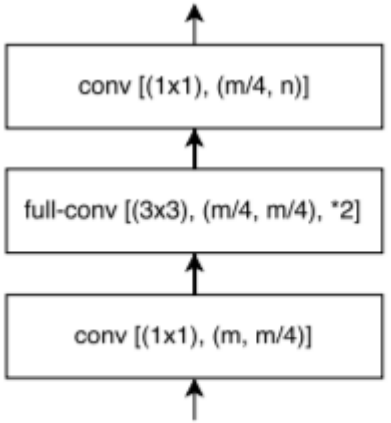
LinkNet34 (April 24)

- LinkNet with ResNet34 Encoder Pre-Trained on ImageNet
- IoU scores on Validation Set： 0.6300





Convolutional modules in *encoder-block (i)*



Convolutional modules in *decoder-block (i)*

Abhishek Chaurasia, Eugenio Culurciello LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation arXiv:1707.03718

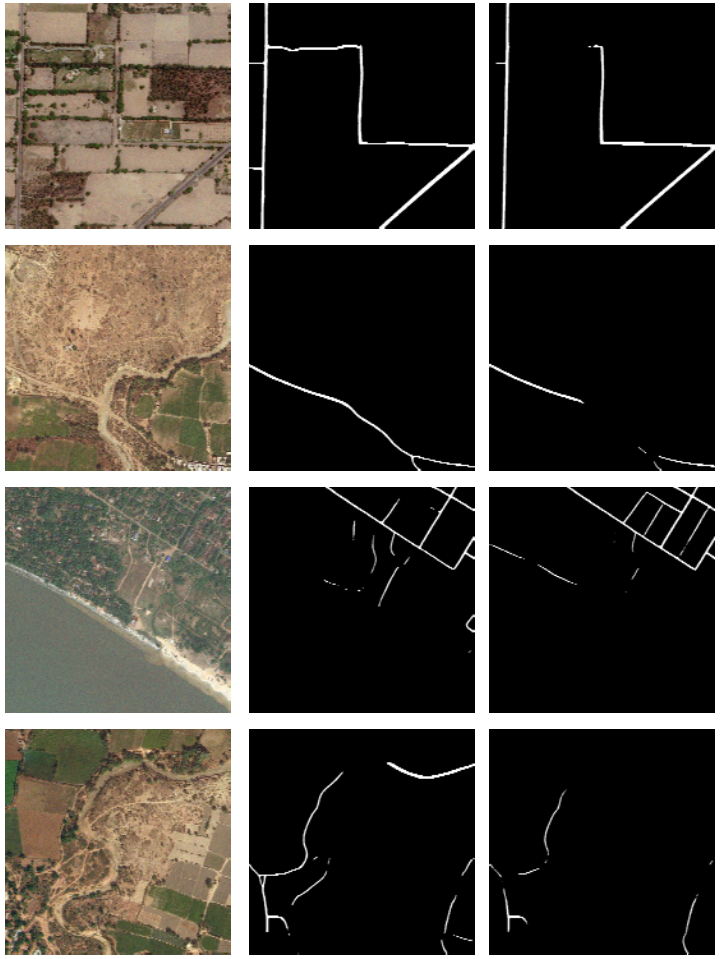
Update the Network

- Unet with TTA&Data Augmentation(0.6294)
- Switch to LinkNet34(0.6300)
- Why is the LinkNet34 just a little bit better?

Review the Predictions

We evaluated the IoU of masks predicted by Unet and masks predicted by LinkNet34, and 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.

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.



Input Unet LinkNet34

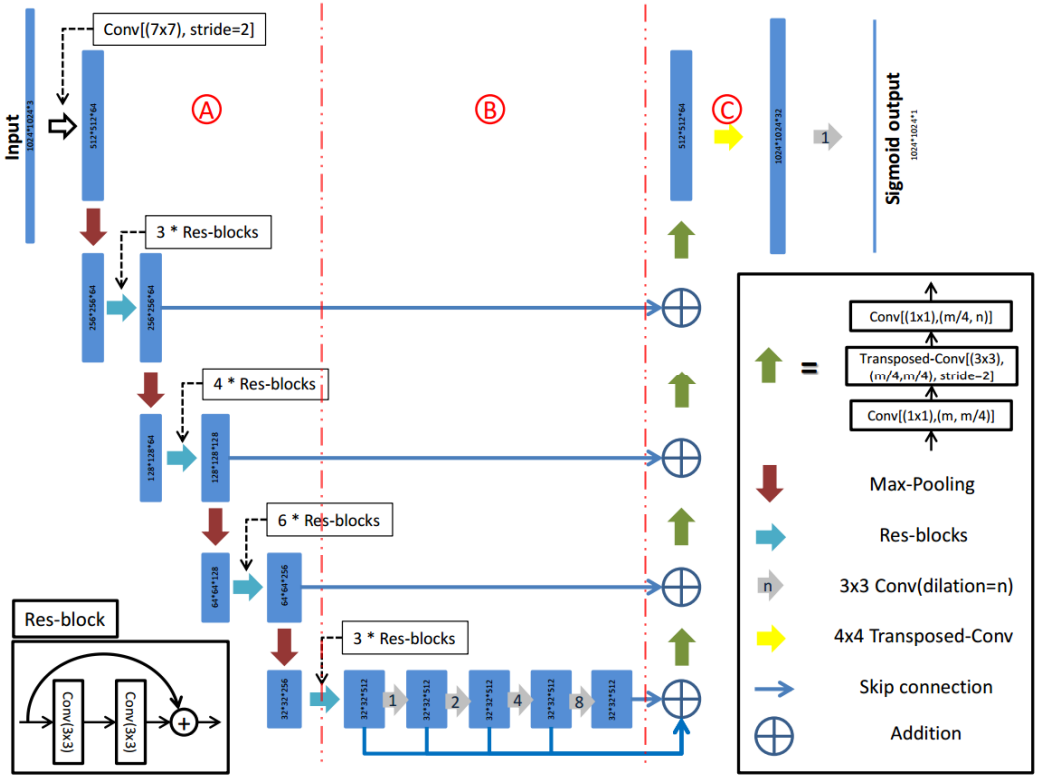
Update LinkNet34 and Unet

- LinkNet34, connectivity problem
  - Preserve the detailed spatial information
  - Extend receptive field
- Unet, judging problem
  - Pretrained

D-LinkNet Architecture

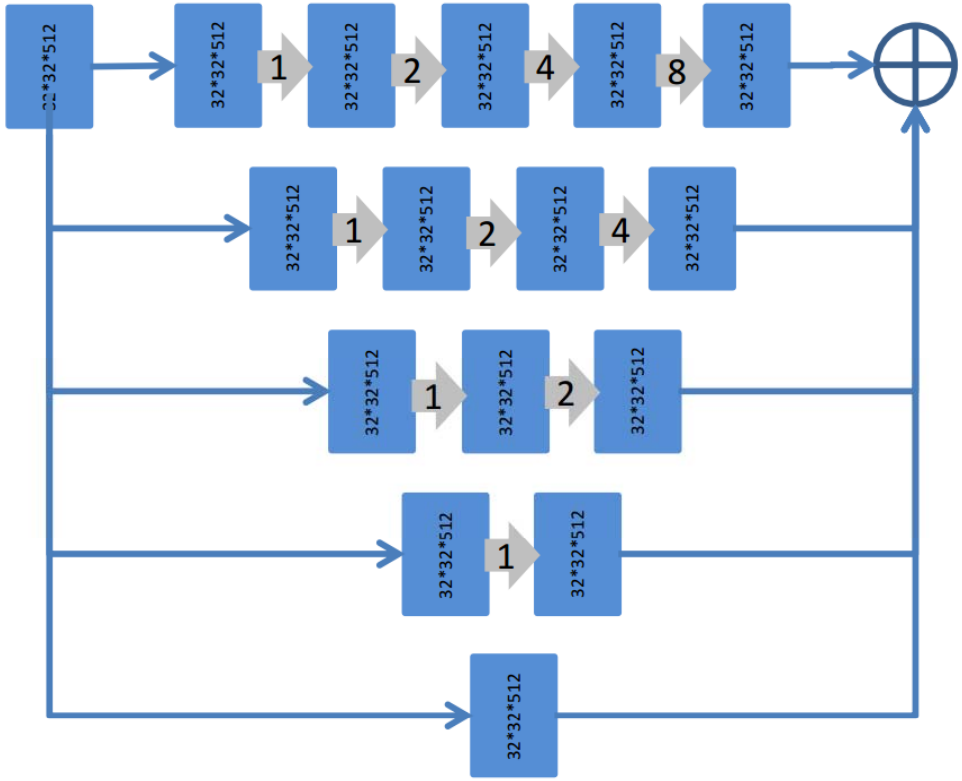
—Update LinkNet34 to D-LinkNet34

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.



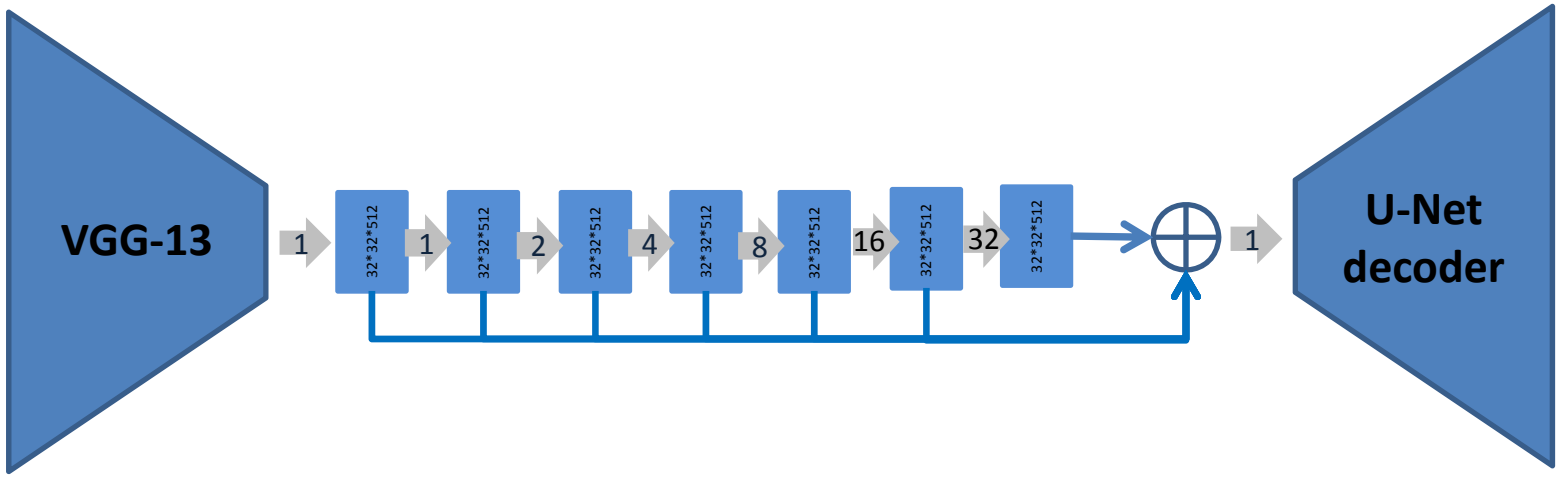
Unroll the Center Part

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.



D-Unet Architecture

—Update Unet to D-Unet



- Pretrained VGG-13 as encoder
- Dilated Convolution layers

Scores on Validation Set

Unet(7 pooling layers, no-pretrain)	0.6294
LinkNet34(pretrained encoder)	0.6300
Ensemble Unet and LinkNet34[1]	0.6394
D-LinkNet34	0.6412
Ensemble D-LinkNet34, Unet and LinkNet34[2]	**0.6466**

[1] 0.5*Unet+0.5*LinkNet34

[2] 0.25*Unet+0.25*LinkNet34+0.5*D-LinkNet34

Final Scores on Test Set

	IoU	Parameter Number	Training Time
D-Unet	0.6194	73M	160H[4]
D-LinkNet34[1]	0.6283	119M	35H[4]
D-LinkNet50[2]	**0.6342**	**831M**	**70H[3]**
D-LinkNet101(Not Converge)	0.6237	904M	>120H[4]
Final Score	**0.6342**		

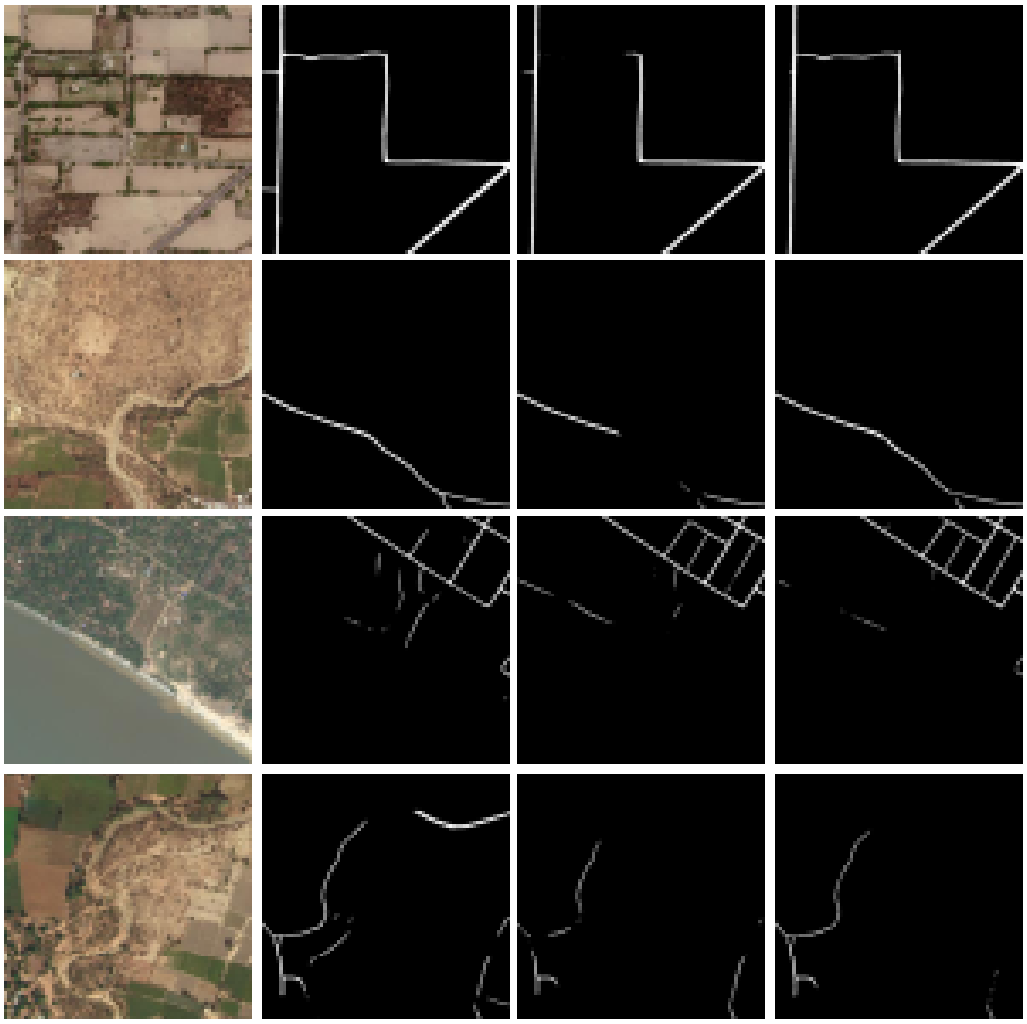
[1] Ensemble Unet, LinkNet34 and D-LinkNet34, , no cross validation, predicting with TTA

[2]Single model, no cross validation, predicting with TTA

[3] Trained on 4 NVIDIA GTX1080 GPUs

[4] Trained on 2 NVIDIA GTX1080 GPUs

Results Visualization



The first two rows are examples showing the road connectivity problem in LinkNet34. There are several road interruptions in LinkNet34 results.

The last two rows are examples showing the false predicting of Unet. Unet is more likely to wrongly recognize roads as background or recognize something non-road like rivers as roads.

DLinkNet avoids weaknesses in Unet and LinkNet34, and makes better predictions.

Feature Work

- D-LinkNet still has the false 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.

PPT&PAPER&CODE

https://github.com/zlkanata/DeepGlobe-Road-Extraction-Challenge

Any Question Email to:
zhoulichen@bupt.edu.cn
wuming@bupt.edu.cn