THE LAND COVER CLASSIFICATION USING A FEATURE PYRAMID NETWORKS ARCHITECTURE FROM SATELLITE IMAGERY

Q. Zhang [1] *, Y. Zhang [1], P. Yang [1], Y. Meng [1], S. Zhuo [1],Z. Yang [1]

[1] The Third Institute of Photogrammetry and Remote Sensing, Ministry of Natural Resources, Chengdu, China-(scrs_qiaozh, newonemylove)@163.com, yapa1228@gmail.com, (mengyin, zsong81)@126.com, yangzy0007@163.com

KEY WORDS: Land cover classification, deep learning, High spatial resolution image, Feature pyramid networks, transfer learning

ABSTRACT:

Extracting land cover information from satellite imagery is of great importance for the task of automated monitoring in various remote sensing applications. Deep convolutional neural networks make this task more feasible, but they are limited by the small dataset of annotated images. In this paper, we present a fully convolutional networks architecture, FPN-VGG, that combines Feature Pyramid Networks and VGG. In order to accomplish the task of land cover classification, we create a land cover dataset of pixel-wise annotated images, and employ a transfer learning step and the variant dice loss function to promote the performance of FPN-VGG. The results indicate that FPN-VGG shows more competence for land cover classification comparing with other state-of-the-art fully convolutional networks. The transfer learning and dice loss function are beneficial to improve the performance of on the small and unbalanced dataset. Our best model on the dataset gets an overall accuracy of 82.9%, an average F1 score of 66.0% and an average IoU of 52.7%.

1. INTRODUCTION

Many global and regional applications require land cover information about Earth's surface. Extracting land cover from satellite imagery is considered as a low cost way and has been applied in many fields such as land resource management, environmental protection. With the development of remote sensing technology, the spatial resolution of satellite images is higher and higher, which provides more information for land cover classification but also brings great challenges (Tong et al., 2018). Thus it is very difficult to find an universal method for land cover classification from the images covering different geographical areas.

The prevalent remote sensing classification methods are mainly based on the spectral and spatial features. These methods consist of two sections: feature extraction and feature classification. Firstly, the features are extracted by manually designed operators such as scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG) et al (Yang, Newsam, 2013). Then the features are classified by classifiers such as support vector machine (SVM), and conditional random field (CRF) et al (Melgani, Bruzzone, 2004, Li et al., 2015). However, these methods is hard to classify the images in complex conditions.

In recent years, deep learning methods have surpassed traditional methods in various computer vision tasks, such as object detection, classification. Convolutional Neural Networks (CNNs) are the most representative deep learning models, which are constructed in deep hierarchical architectures and capable of extracting the intrinsic features of data. In 2012, Professor Hinton and his student Alex (Krizhevsky et al., 2012) won the ILSVR (ImageNet Large Scale Visual Recognition Competition) by employing CNNs. After that, the deep learning method has been widely used in remote sensing (Zhu et al., 2017) and other fields. At first, remote sensing scientists exploited the deep learning in the scene classification which is a more coarse classification method than pixel level (Nogueira et al., 2016, Zhong et al., 2016, Tong et al., 2018). The Fully Convolutional Networks (FCN) (Long et al., 2015), which replaces the fully connected layers with convolution layers, could directly obtain the pixel-wise classification results (Wu et al., 2018, Zhang et al., 2018). There are several FCN networks such as FCN-8s (Long et al., 2015), Segnet (Badrinarayanan et al., 2015) and U-net (Ronneberger et al., 2015).

Although FCNs are the most popular approach to pixel-wise classification, they require huge computing resources, as well as a large dataset of pixel-wise annotated images, which impedes their application in remote sensing. There are very few pixel-wise annotated land cover dataset such as ISPRS Benchmark dataset. In order to meet our classification from satellite imagery, we create a land cover dataset consisting of images and manually pixel-wise annotated labels. We design a FCN architecture which combines the Feature Pyramid Networks (Lin et al., 2017) and VGG (Simonyan, Zisserman, 2015), and overcome the limitation of small and unbalanced dataset by using a transfer learning step and the variant dice loss function.

2. METHODOLOGY

2.1 Network architecture

Feature Pyramid Networks (FPN) is coined to detect objects (Lin et al., 2017). FPN builds feature pyramids inside convolutional networks which are critical to address multiscale problems. We present a new network architecture named FPN-VGG (Figure 1) which is modified from FPN and combine the VGG network. The VGG network, proposed by the Visual Geometry Group from University of Oxford (Simonyan, Zisserman, 2015), is used as a feature extractor for FPN.

* Corresponding author

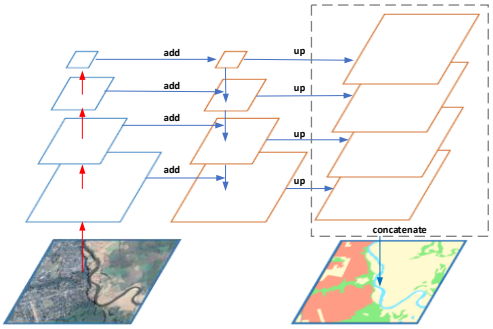


Figure 1. FPN architecture for land cover classification from satellite imagery. The decoder part (shown in orange) of FPN consists of a feature pyramid with 4 pyramid levels, and each pyramid level is up-sampled to generate layers with size of 1/4 input image size. And then they are assembled and up-sampled to generate segmentation mask.

To train a deep learning network, the following problems always impede us to obtain the best model: (a) The overfitting led by the small training dataset. (b) Slow convergence because of random initialization. In order to overcome these two problems, we employ a transfer learning strategy by initializing the feature extractor of FPN-VGG with a VGG16 pretrained model. Transfer learning has been proved a good way to train deep neural networks on small dataset (Huh et al., 2016). It has been proved that ImageNet pretrained networks could promote the performance of classification on remote sensing data (Marmanis et al., 2016). Thus, we transfer the parameters of VGG16 model (excluding the top fully connected layers) trained on 2012 ImageNet dataset to initialize the encoder part of FPN-VGG (Figure 2).

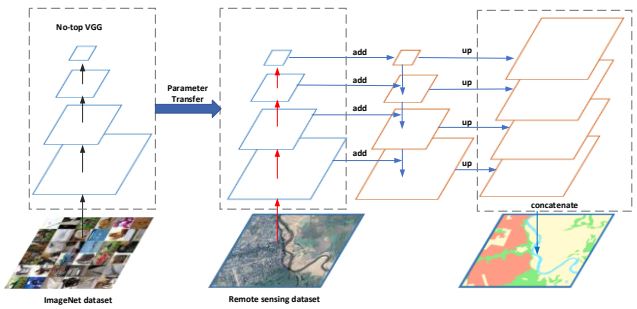


Figure 2. FPN-VGG architecture

2.2 Loss function and accuracy assessment

In multiclass classification task, categorical cross entropy loss (L_ categorical_crossentropy) is the most commonly used loss function. It is calculated from categorical cross entropy between the ground truth (gt) and the prediction (pr).

$$L\_\text{categorical}\_\text{crossentropy}=-\text{gt}\cdot\log(\text{pr}) \quad (1)$$

We have also employed the dice loss (L_dice) to train the FPN-VGG network. The dice loss is defined as follow,

$$L\_dice=1-F \quad (2)$$

Where,

$$F\_\beta=(1+\beta^2)\frac{precision\cdot recall}{\beta^2\cdot precision+recall} \quad (3)$$

Where β is a coefficient for precision and recall balance, and it is set to 1 in our work.

We also used the sum of L_categorical_crossentropy and L_dice as a variant dice loss function. This loss function is named L_cce_dice.

$$L\_cce\_dice=L\_\text{categorical}\_\text{crossentropy}+L\_\text{dice} \quad (4)$$

We employ F1 score and IoU (intersection over union), which are the most common indexes used to assess the accuracy for semantic segmentation (Maggiori et al., 2017), to assess the classification accuracy of remote sensing images. For F1, the relative contribution of precision and recall are equal.

$$F1=\frac{2\cdot precision\cdot recall}{precision+recall} \quad (5)$$

Where,

$$precison=\frac{TP}{TP+FP},\quad recall=\frac{TP}{TP+FN} \quad (6)$$

$$IoU=\frac{TP}{TP+FP+FN} \quad (7)$$

Where, TP denotes true positives, FP denotes false positives, FN denotes false negatives. F1 and IoU reach the best value at 1 and worst score at 0.

For multiclass classification task, we employ mF1 and mIoU to assess the accuracy for remote sensing image classification.

$$mF1=\frac{1}{n}\sum_{i-1}^{n}F1(i) \quad (8)$$

$$mIoU=\frac{1}{n}\sum_{i-1}^{n}IoU(i) \quad (9)$$

3. RESULTS AND DISCUSSION

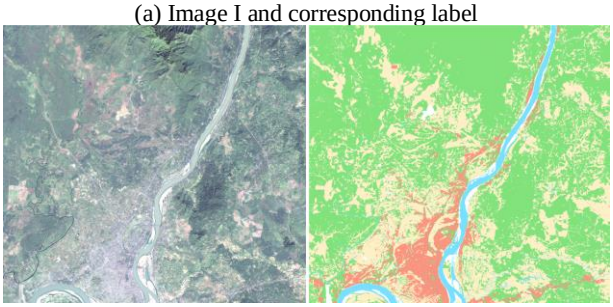
In this section, we present a land cover dataset and design experiments to analyse the performance of FPN-VGG.

3.1 Dataset and Experiments

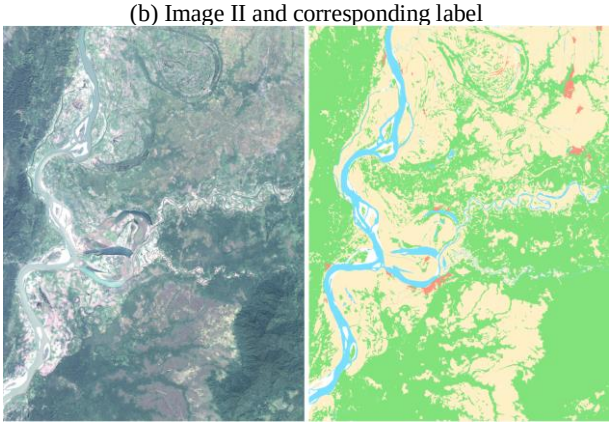
3.1.1 Dataset: We prepare a dataset of land cover classification from high resolution satellite images. The dataset consists of a set of Digital Orthophoto Maps (DOM) and the corresponding annotated labels. The DOM achieved from ZY-3 satellite consists of the four spectral bands in the visible (VIS: red(R), green(G), blue(B)) and in the near infrared (N).

We have extracted 3 images from large DOM as training data which are manually annotated with 6 classes (background, low vegetation (lowVeg), tree, building, road, water). These classes are commonly used in the applications of land cover. These images cover 3400 km^2 on East Asia with 5.8m ground resolution. The images and labels are shown in Figure 3. The pixel number of each class is shown in Figure 4.

(a) Image I and corresponding label



(b) Image II and corresponding label



(c) Image III and corresponding label

Figure 3. The images and annotated labels in the dataset

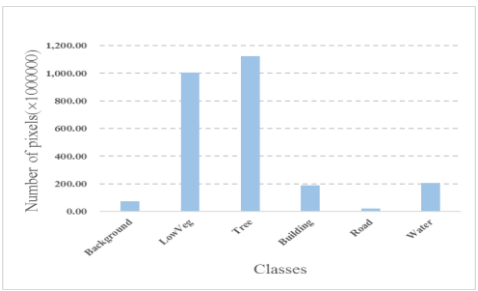
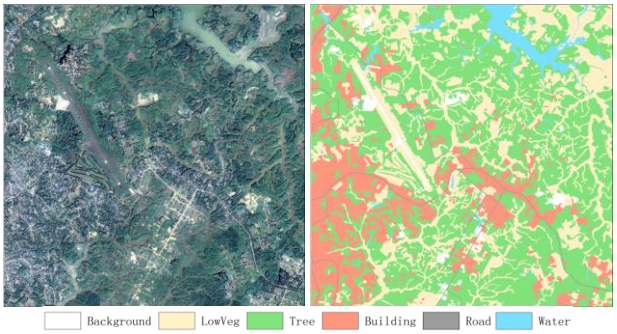


Figure 4. The pixel number of each class in the dataset

It is known that working with large image patches could maximize the advantage of CNN (Wu et al., 2018). However, the maximum of patch size is limited by the memory of the GPU hardware. Thus we have created training data by extracting image patches of size 480×480. After slicing the original images and labels, we split the 480×480 patches into a training set and a validation set with a ratio of 0.25. In order to test the performance of models, a test image has cropped from a large DOM and manually annotated it as a reference map (Figure 5). The test image is also from ZY-3 satellite on East Asia covering 40 km^2.



(a) Test image (b) Reference map

Figure 5. The test image and corresponding reference map

3.1.2 Design of experiments: In order to test the performance of FPN-VGG on land cover classification task, we design the following set of experiments: (a) Training the FPN-VGG and other state-of-the-art FCN networks, and comparing their performances. (b) Training and testing FPN-VGG with different input spectral bands. (c) Training FPN-VGG with ImageNet pretrained model. (d) Training and testing FPN-VGG with dice loss, and comparing the performances of different loss function. All above experiments are carried out on the same computer with a NVIDIA GeForce GTX1080TI GPU.

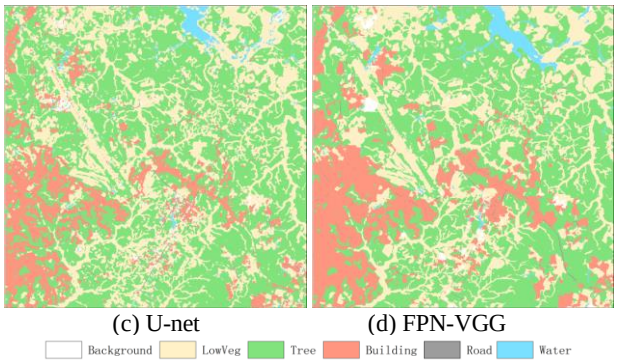
3.2 Comparison with other networks

The performance and classified maps of FPN-VGG and other FCN networks are shown in Table 1 and Figure 6. These networks are all trained and tested with input data of RGB bands.

From Table 1, the worst result for all models comes in the class “road”. And the roads cannot be extracted by Segnet. This problem is mainly owed to the fact that proportion of road class in training dataset is much smaller than other classes (see Figure 4). Overshadowed by contiguous trees and buildings might also give rise to this problem (see Figure 5(a)).

From the overall performance, the FPN-VGG surpasses the other three networks. The FPN-VGG takes the highest mF1 (66.0%), mIoU (52.7%) and OA (82.9%), followed by the FCN-8s. The performances of Segnet and U-net are worse than FCN-8s. From Figure 6, the map classified by U-net is more fragmented than maps classified by other networks.

(a) FCN-8s (b) Segnet



(c) U-net (d) FPN-VGG

Figure 6. The predictions by different networks with input data of RGB images

3.3 Performance of FPN-VGG models with different input bands

Being different from natural photos, the satellite images always have more than 3 spectral bands containing the visible (RGB) and the near infrared (N). Thus, we train the FPN-VGG by supplementing the N band.

The accuracy of FPN-VGG models with different input spectral bands is listed in Table 2. The performance of the model with NRG input is significantly better than that of model with RGB input. Although the OA of model with NRG input is only 1.0% higher than that of model with RGB input, the mF1 and mIoU are 4.7% and 5.4% higher respectively. This improvement is mainly contributed by the improvement of "road" and "water". This indicates that the NRG bands has the advantage for extracting road and water.

While the model with 4 bands input has one more band than model with NRG input, the performance with 4 bands input is slightly lower. This means that more input bands may not obtain better performance. We should choose the optimal band combination for the specific classification task.

3.4 Transfer learning

In this section, we initialize the FPN-VGG with ImageNet pretrained model and NRG input bands in this section. In Table 3 we provide the comparative results of ImageNet initialization and random initialization. Comparing to the result of random initialization, the FPN-VGG initialized by ImageNet pretrained model shows the improvement on all classes except background. The mF1 , mIoU and OA are improved by 1.6%, 1.9% and 1.0% respectively. This indicates that the ImageNet model is beneficial to improve the model with NRG input bands despite it is trained on RGB ImageNet dataset.

3.5 Training with dice loss function

In order to exam the classification ability of FPN-VGG with dice loss on unbalanced dataset, we present the results of models with different loss function are shown in Table 4.

Comparing to the model with *L*_categorical_crossentropy, although the OA, mF1 and mIoU of model with *L*_dice is worse, the F1 of "background" and "road" is 2.7% and 2.9% higher. It shows that the dice loss could improve the classification ability of minority classes.

Furthermore, the model with *L*_cce_dice has the highest mF1 (75.9%) and mIoU (63.2%). The best prediction map is shown in Figure 7. Comparing to model with *L*_categorical_crossentropy, the mF1 and mIoU of *L*_cce_dice have been improved 3.6% and 3.2%. Especially, the F1s of "background", "building", "road" and "water" classified by the model with *L*_cce_dice are higher than those of other two models. It can be seen that the *L*_cce_dice is more competent for unbalanced multiclass classification than *L*_categorical_crossentropy and *L*_dice.

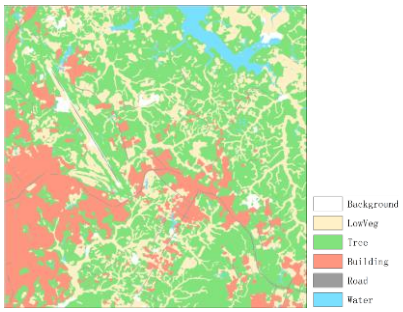


Figure 7. The best prediction by FPN-VGG model

4. CONCLUSION

In this paper we present a new deep learning modelling framework for land cover classification of high spatial resolution satellite imagery. The framework is named FPN-VGG which is based on feature pyramid networks combining with VGG16. The performance of our framework is evaluated on a dataset manually annotated. The dataset consists images with four spectral bands (Blue, Green, Red and the Near infrared) and corresponding labels of 6 classes (background, low vegetation, tree, building, road, water). The training dataset are extracted from ZY-3 satellite covering on East Asia.

We found that the FPN-VGG could extract more accurate land cover map from satellite images than other state-of-the-art fully convolutional networks (FCN-8s, Segnet, U-net). The experiment results show that inputting with NRG spectral bands and initializing by ImageNet model could improve the performance of FPN-VGG on small dataset. In addition, the model trained with the sum of the categorical cross entropy loss and dice loss is more competent for classification on unbalanced dataset.

Networks	Background	LowVeg	Tree	Building	Road	Water	mF1 (%)	mIoU (%)	OA (%)
FCN-8s	**54.7**	74.3	86.5	81.6	22.3	63.1	63.8	50.1	80.9
Segnet	24.7	72.5	85.3	79.7	0.0	71.2	55.6	44.5	79.4

U-net	26.6	73.9	86.3	77.1	21.0	55.6	56.8	43.8	79.2
FPN-VGG	44.9	**77.5**	**87.6**	**83.5**	**31.1**	**71.6**	**66.0**	**52.7**	**82.9**

Table 1. Performance of different networks with input data of RGB images. The metric for individual class is F1 (%).

Spectral bands	Background	LowVeg	Tree	Building	Road	Water	mF1 (%)	mIoU (%)	OA (%)
RGB	44.9	77.5	87.6	83.5	31.1	71.6	66.0	52.7	82.9
NRG	**48.5**	**79.2**	88.1	83.1	**37.2**	**88.0**	**70.7**	**58.1**	**83.9**
4 Bands	46.5	77.7	**88.2**	**83.6**	34.4	85.3	69.3	56.8	83.5

Table 2. Performance of FPN-VGG models with different input bands. The metric for individual class is F1 (%).

Initialization approach	Background	LowVeg	Tree	Building	Road	Water	mF1 (%)	mIoU (%)	OA (%)
Random	**48.5**	79.2	88.1	83.1	37.2	88.0	70.7	58.1	83.9
ImageNet	48.0	**80.7**	**88.7**	**84.4**	**42.1**	**89.6**	**72.3**	**60.0**	**84.9**

Table 3. Performance of FPN-VGG models trained with different initialization approaches. The metric for individual class is F1 (%).

Loss function	Background	LowVeg	Tree	Building	Road	Water	mF1 (%)	mIoU (%)	OA (%)
Categorical Cross entropy	48.0	**80.7**	**88.7**	84.4	42.1	89.6	72.3	60.0	**84.9**
Dice	50.7	74.6	85.5	80.9	45.0	83.6	70.1	56.2	80.8
Cce_dice	**61.6**	78.3	87.9	**84.7**	**51.2**	**91.4**	**75.9**	**63.2**	84.1

Table 4. Performance of FPN-VGG models trained with different loss functions. The metric for individual class is F1 (%).

REFERENCES

Badrinarayanan, V., A. Kendall and R. Cipolla, 2015: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12): 2481-2495.

Huh, M., P. Agrawal and A. A. Efros, 2016: What makes imagenet good for transfer learning? *arXiv:1608.08614*.

Krizhevsky, A., I. Sutskever and G. E. Hinton, 2012: ImageNet Classification with Deep Convolutional Neural Networks. *International Conference on Neural Information Processing Systems*, 25**:** 1097-1105.

Li, E., J. Femiani, S. Xu, X. Zhang and P. Wonka, 2015: Robust rooftop extraction from visible band images using higher order CRF. *IEEE Transactions on Geoscience and Remote Sensing*, 53(8): 4483-4495.

Lin, s.-Y., P. Doll´ar, R. Girshick, K. He, B. Hariharan and e. Belongie, 2017: Feature Pyramid Networks for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition*.

Long, J., E. Shelhamer and T. Darrell, 2015: Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition. Boston*, MA, USA, IEEE, 7298965.

Maggiori, E., Y. Tarabalka, G. Charpiat and P. Alliez, 2017: Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55: 645-657.

Marmanis, D., M. Datcu, T. Esch and U. Stilla, 2016: Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1): 105-109.

Melgani, F. and L. Bruzzone, 2004: Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8): 1778-1790.

Nogueira, K., O. a. A. B. Penatti and J. A. d. Santos, 2016: Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classication. *Pattern Recognition*, 61(7): 539-556.

Ronneberger, O., P. Fischer and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, Springer**:** 234-241.

Simonyan, K. and A. Zisserman, 2015: Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Tong, X.-Y., Q. Lu, G.-S. Xia and L. Zhang, 2018: Large-scale land cover classification in GaoFen-2 satellite imagery. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Valencia, Spain.

Wu, G., X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu and R. Shibasaki, 2018: Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing*, 10(3): 407-424.

Yang, Y. and S. Newsam, 2013: Geographic image retrieval using local invariant features. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2): 818-832.

Zhang, W., H. Huang, M. Schmitz, X. Sun, H. Wang and H. Mayer, 2018: Effective fusion of multi-modal remote sensing

data in a fully convolutional network for semantic labeling. *Remote Sensing*, 10(52): 1-14.

Zhong, Y., F. Fei and L. Zhang, 2016: Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing*, 10(2): 025006-025001-025020.

Zhu, X. X., D. Tuia, L. Mou, G.-S. Xia, L. Zhang and F. Xu, 2017: Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4): 8-36.