Semantic Segmentation of Multispectral Images using Res-Seg-net Model

Nidhi Saxena
Department of Computer Science and Engineering,
Indian Institute of Technology, Roorkee, India
Email: nidhi.pdfcse@iitr.ac.in

N. Kishore Babu
Department of Computer Science and Engineering,
Indian Institute of Technology, Roorkee, India
Email:kbabu89@cs.iitr.ac.in

Balasubramanian Raman
Department of Computer Science and Engineering,
Indian Institute of Technology, Roorkee, India
Email: balarfcs@iitr.ac.in

***Abstract*—Semantic segmentation is pixel-wise labeling of the image. Recently deep convolutional neural network (DCNN) providing progressive results in semantic segmentation. However, in remote sensing multispectral imagery very limited work has been done due to lack of training dataset. In this paper, a Res-Seg-net model is proposed for the semantic segmentation which is motivated by the existing Resnet and Segnet models. This model consists of encoder-decoder parts in which residual mapping is followed. For validation and testing of the proposed model, the RIT-18 dataset of multispectral imagery is used. The comparison results of the experiment on a multispectral imagery dataset have demonstrated the effectiveness of the proposed model.**

***Index Terms*—Convolutional neural network, Multispectral images, Semantic Segmentation.**

I. Introduction

Segmentation of an image is producing labels for each pixel based on different categories with more applications ranging from low-level to high-level applications. It has been a challenging task to locate an image with different background, visual features and conditions. Recently it has become more popular in the machine learning areas of research such as feature processing and deep learning techniques. In this process, the input data are images and learning the representation of those images led to high progress in various fields of research such as edge detection [1], object detection [2], speech recognition [3] and image classification [4]. The effort of deep learning techniques achieves high-level abstraction of data at various levels based on the cognitive process of the human brain, which typically starts from one level to a higher level of abstraction. Learning can be done with different types of architectures like Deep belief networks, auto association networks, and convolutional networks. Among these, deep convolutional networks are highly accurate and high capacity learning models with more number of parameters, which are optimized by using trained examples.

Recently there is rapid development not only in data collection, which is remotely sensed but also in the transfer of the data and storage of the data. Due to the availability of resources, mixed land covers with different types for a single pixel to be used. Most of the remote sensing imagery classification models use characteristics of land cover like spatial, structural information and shapes. These cause wastage large amounts of resources and creating a problem for remote sensing techniques. An increase in different varieties of satellites and different types of sensors is leading to getting better imagery with high spatial resolution. Therefore, an efficient technique by using which extraction of useful information from high spatial resolution remote sensing imagery plays a vital role in current research. There are many deep learning models for multispectral imagery in order to do segmentation semantically. These models suffer from some semantic segmentation problems like overfitting problems, band limitation in the input image and more parameter requirement in the CNN model. To the reduction of all these problems in this paper, a Res-Seg-net model is proposed. The Res-Seg-net model is a convolutional neural network-based encoder-decoder model. It trained the network by available multispectral ground truth labeled images, therefore, it provides the segmented images with no overfitting. This model design for six bands of input images. This model has more effective than the existing models is obtained by the experiments. In which existing schemes of semantic segmentation based on Super vector machine (SVM), Segnet model without transfer learning (SegNet-RI), Segnet model with transfer learning (SegNet-TL), Segnet model with transfer learning and principal component analysis (SegNet-TL-PCA), k-nearest neighbor (k-NN), stacked convolutional autoencoder (SCAE) [5], [6] are used for the comparison. To validation and testing of schemes RIT-18 dataset of multispectral imagery is used. The result of the experiment shows that the proposed model gives better results.

The rest of this paper is organized as follows. Section II reviews the related work on semantic segmentation. Section III proposes a semantic segmentation model. Section IV shows the experimental results, and finally, Section V concludes the paper.

II. Review on semantic segmentation

In recent years, semantic pixelwise segmentation is a progressive subject of research in many areas. It is applied on challenging datasets [7], [8], [9], [10]. All recent approaches have aimed to produce high-quality results by trying to predict labels for all pixels by designing high-efficiency models and training datasets. The semantic segmentation of remote sensing

978-1-7281-6332-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICSC.2020.00030

IEEE computer society

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on March 16,2023 at 20:07:41 UTC from IEEE Xplore. Restrictions apply.

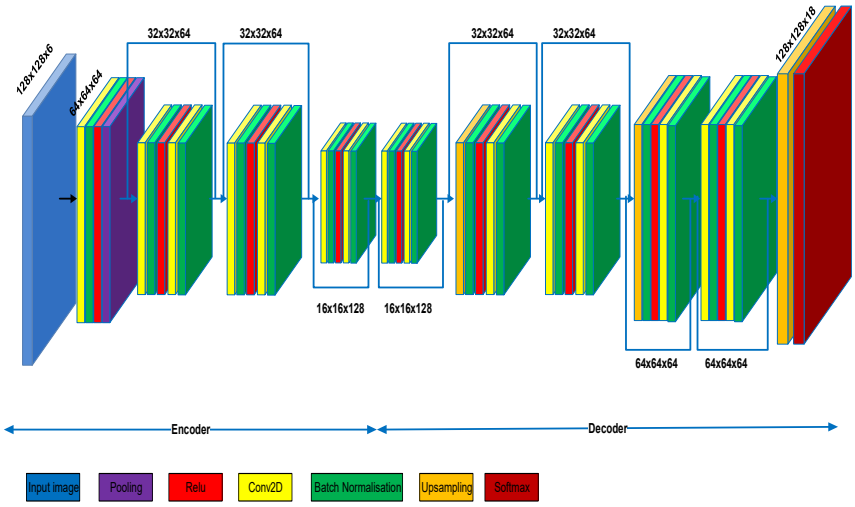
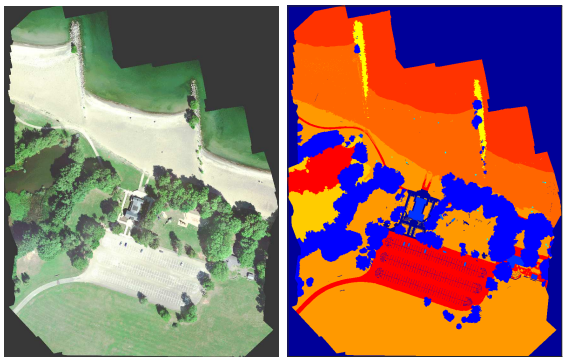


Fig. 1: Block diagram of the proposed Res-Seg-net model

imagery provides the end-user pixel-wise classification map for a given scene. In the semantic pixel-wise segmentation, deep convolution neural network (DCNN) is used and it is achieved significant results by the progress of DCNN models [11] trained with large quantities of labeled imagery. In the multispectral imagery, the available quantity of labeled data is minuscule therefore DCNNs less successful for remote sensing applications. In the existing schemes, semantic segmentation is applied on only RGB band images therefore extra information given by the other bands is missed in the segmentation [12], [13]. In the semantic segmentation, the fully connected layer maintains a high-resolution feature map in the deepest CNN model which requires more parameters [14]. In this paper, to reduce the overfitting problem, band limitation and more parameter requirement in the semantic segmentation, a new Res-Seg-net model is proposed. As a result, it reduces the number of parameters without disturbing the performance of the model by which it reduces memory consumption and inference time.



(a) (b)

Fig. 2: Input image dataset of RIT-18 (a) input image, (b) Ground truth labeled image

III. Proposed Semantic Segmentation

A. Res-Seg-net model

The proposed Res-Seg-net model is inspired by the existing Resnet and Segnet models. The Resnet model works on residual mapping. At the extreme case, if an identity mapping is optimal, it would be easy to push the residuals to zero, so that an identity mapping could be fitted by the stack of nonlinear layers. In the Segnet model [11], non-linear layers are connected to generate an auto-encoder. At the output of the autoencoder model layer, a fully connected layer is discarded to retaining higher resolution feature maps. In results, it achieved similar performance with low memory consumption and inference time.

These all the benefits of Resnet and Segnet models have been used in the proposed (Res-Seg-net) model. The proposed Res-Seg-net model shown in fig. 1 have both encoder and decoder networks to followed pixel-wise classification. An encoder part consists of convolution, Relu, batch-normalization, max-pooling layers and in the decoder part upsampling, convolution, Relu, batch-normalization, softmax layers are stacked to obtain the abilities for each pixel independently.

1) Encoder: The Input image H is applied on the CNN model for obtaining m intermediate features using W convolution filters $F^{(1)} = \{F_1^{(1)}, F_2^{(1)}, ..., F_W^{(1)}\}$ by following equation

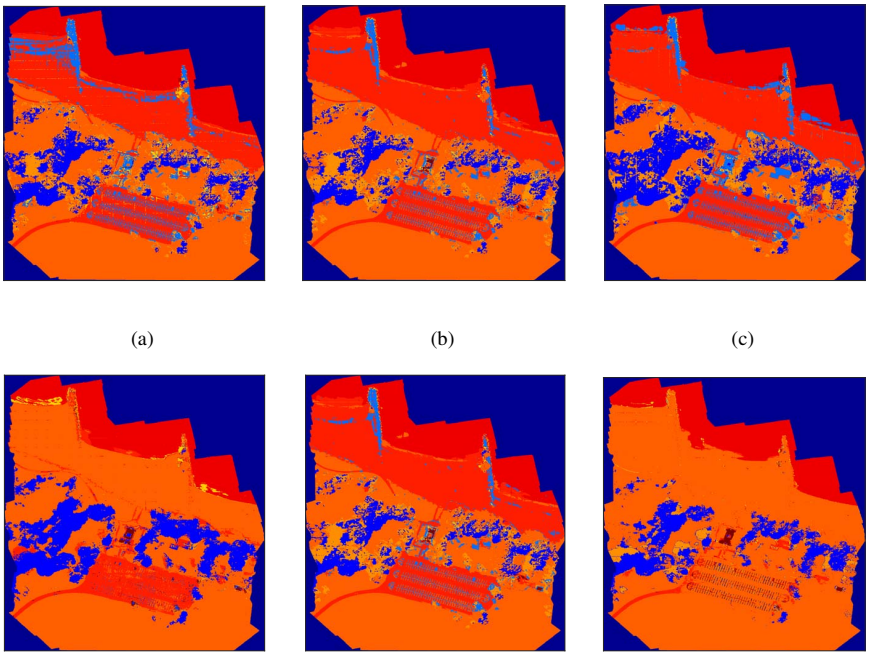
$$T_m = s(H * F_m^{(1)} + b_m^{(1)}),\ m = 1, 2, ..., W, \qquad (1)$$

where s and b are the sigmoid activation function and bias vectors for m^{th} feature map.

2) Decoder: In this part, m intermediate features are reconsturcted (G) from encoded T_m features of input image H and convolution filter $F^{(2)} = \{F_1^{(2)}, F_2^{(2)}, ..., F_W^{(2)}\}$ expressed by following equation

$$G = s(T * F_m^{(2)} + b_m^{(2)}),\ m = 1, 2, ..., W, \qquad (2)$$

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on March 16,2023 at 20:07:41 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c)

(d) (e) (f)

Fig. 3: Comparison of Proposed Res-Seg-net and U-net models at different epochs (a) proposed Res-Seg-net-100 model at 100 epochs, (b) proposed Res-Seg-net-75 model at 75 epochs, (c) proposed Res-Seg-net-50 model at 50 epochs, (d) U-net-100 model at 100 epochs, (e) U-net-75 model at 75 epochs, (f) U-net-50 model at 50 epochs,

where W and b represents convolution filter and bias vectors, respectively, with stochastic gradient descent (SGD) based optimization technique and backpropagation (BP). All the parameters {W,b} in the network can be iteratively learned to reach an optimal allocation to reduce the prediction loss between network output G and ground truth H. In the encoder-decoder parts, 15 convolution layers are used to segment multispectral images for object detection.

IV. Experiment Result

A. Data

To validation and testing of proposed model, RIT-18 [6], [15] multispectral dataset is used. This dataset has six multi-spectral bands covered visible and near-infrared regions. This dataset divided into training and testing, wherein the size of training data is 9394×5642 and testing data 8833×6918.

The CNN models for the proposed scheme were trained and tested on GPU (Nvidia GTX 1080 Ti 11GB with CUDA 10.1) through MATALB 2019a using deep learning toolbox in Windows 10 operating system.

B. Hyperparameters

In the training process of CNN Network, some training hyperparameters are used to make a better-trained model like epochs, epoch intervals, initial learning rate, mini-batch size, image patches and convolution layers (L). In the proposed Res-Seg-net model, training hyperparameters: epochs, initial learning rate, mini-batch size, Regularization coefficient are selected and the values of these hyperparameters are 100, 0.05, 16, 0.0001, respectively and image patches are 128×128. Using these hyperparameter settings, the proposed Res-Seg-net and U-net [15] models have been compared at the different epochs.

C. Results and Discussion

Input images and labeled ground truth images are shown in fig. 2 (a) and (b). In fig. 3, the Res-Seg-net model and the U-net model are compared at 50,75 and 100 epochs. Compared to U-Net, it has been observed that the proposed semantic segmentation for 100 epochs gives better results. In Table II, accuracy, and covered vegetation parts are given for these models at different epochs. It is seen that more accuracy and maximum cover vegetation parts are obtained by the proposed model.

In the experiment, the results of the proposed and existing semantic segmentation schemes are compared. In which existing schemes of semantic segmentation are Segnet model without transfer learning (SegNet-RI), Segnet model with transfer learning (SegNet-TL), Segnet model with transfer learning and principal component analysis (SegNet-TL-PCA), Super vector machine (SVM), k-nearest neighbor (k-NN), stacked convolutional autoencoder (SCAE) [5], [6]. For comparison with the existing scheme used training hyperparameters epochs, initial

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on March 16,2023 at 20:07:41 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Comparision of accuracy of proposed and existing semantic segmentation schemes using RIT-18 dataset

Class	SVM	k-NN	SCAE	SegNet-RI	SegNet-TL	SegNet-TL-PCA	CoinNet	Res-Seg-net
RoadMarkings	51.0	65.0	37.0	71.0	63.1	21.7	85.1	78.0
Tree	43.5	71.0	62.0	79.2	77.7	71.7	77.6	79.0
Building	1.5	0.3	11.1	42	60.1	52.8	52.3	0.0
Vehicle	0.2	15.8	11.8	0.0	4.7	59.1	59.8	49.4
Person	19.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Lifeguard Chair	22.9	1.0	29.4	0.0	0.0	0.0	0.0	0.0
Picnic Table	0.8	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Orange Pad	15.2	14.6	82.6	0.0	0.0	0.0	0.0	0.0
Buoy	0.7	3.6	7.2	0.0	0.0	0.0	0.0	0.0
Rocks	20.8	34.0	36.0	76.3	96.9	89.3	84.8	84.1
Low Vegetation	0.4	2.3	1.1	1.0	12.5	1.3	4.1	12.1
Grass/Lawn	71.0	79.2	84.7	98.0	93.0	96.1	96.7	96.5
Sand/Beach	89.5	56.1	85.3	6.4	73.9	77.6	92.1	92.2
Water/Lake	94.3	83.6	97.5	90.4	93.9	98.4	98.4	76.4
Water/Pond	0.0	0.0	0.0	1.7	0.5	87.6	92.7	69.0
Asphalt	82.7	80.0	59.8	73.1	42.4	61.2	90.4	54.3

TABLE II: Comparison of proposed and U-net models based semantic segmentation at different epochs

Models	Epochs	Accuracy	Vegetation cover (%)
Res-Seg-net-100	100	69.1	54.98
Res-Seg-net-75	75	68.0	55.94
Res-Seg-net-50	50	70.0	55.08
U-net-100	100	53.0	84.79
U-net-75	75	53.0	82.27
U-net-50	50	68.7	55.94

learning rate, mini-batch size, Regularization coefficient are 15, 0.01, 12, 0.0001, respectively and the size of image patches is 288×288. In Table I, all the 16 classes are classified in terms of accuracy for proposed and existing schemes. The accuracy of Tree, Rocks, Low vegetation, Sand, Grass classes have some improvement as compared to the existing schemes. Remaining classes are very limited in samples so it does not give improvement. Semantic segmentation by the Res-Seg-net model gives better performance as compared to other existing schemes.

V. Conclusion

In this paper, a Res-Seg-net model based on CNN is presented for semantic segmentation that is inspired by existing Resnet and Segnet models. It consists of encoder-decoder portions in which residual mapping is followed and the fully-connected layer in the decoder output is removed to reduce the parameters in the model. For validation and testing, the RIT-18 dataset of multispectral imagery has been used in which 16 classes exist. The comparison results of the experiment on a multispectral imagery dataset have demonstrated the effectiveness of the proposed model.

References

[1] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3000–3009.

[2] W. Guan, T. Wang, J. Qi, L. Zhang, and H. Lu, "Edge-aware convolution neural network based salient object detection," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 114–118, 2018.

[3] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125 868–125 881, 2019.

[4] J. Jiang, X. Feng, F. Liu, Y. Xu, and H. Huang, "Multi-spectral rgb-nir image classification using double-channel cnn," *IEEE Access*, vol. 7, pp. 20 607–20 613, 2019.

[5] B. Pan, Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu, "Coinnet: Copy initialization network for multispectral imagery semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 816–820, 2018.

[6] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 60–77, 2018.

[7] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[8] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[9] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[12] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5585–5599, 2017.

[13] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474–478, 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] R. Kemker, C. Salvaggio, and C. Kanan, "High-resolution multispectral dataset for semantic segmentation," *arXiv preprint arXiv:1703.01918*, 2017.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on March 16,2023 at 20:07:41 UTC from IEEE Xplore. Restrictions apply.