The Effect of Resnet Model as Feature Extractor Network to Performance of DeepLabV3 Model for Semantic Satellite Image Segmentation

2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS) | 978-1-6654-1464-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/AGERS51788.2020.9452768

Yaya Heryadi
Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta
11480, Indonesia
yayaheryadi@binus.edu

Edy Irwansyah
School of Computer Science, Bina Nusantara University, Jakarta
11480, Indonesia
eirwansyah @binus.edu

Eka Miranda
Information Systems Department Bina Nusantara University, Jakarta
11480, Indonesia
ekamiranda @binus.ac.id

Haryono Soeparno
Bioinformatics and Data Science Research Center Bina Nusantara University, Jakarta
11480, Indonesia
haryono@binus.edu

Herlawati
Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta
11480, Indonesia
mrs.herlawati@gmail.com

Kiyota Hashimoto
Prince of Songkla University, Phuket Campus, Phuket 83120, Thailand
kiyota.h@phuket.psu.ac.th

***Abstract*— Semantic image segmentation is an interesting problem in Computer Vision with many potential applications. The DeepLab model is combined with two other networks: Resnet and Conditional Random Field networks, making the DeepLab model a fairly deep network structure to increase semantic segmentation performance. Many previous studies argued that there are some limits on the deep learning model's depth as the deep structure may lead to vanishing/exploding gradient, which the model's performance. This paper presents an experimental study to compare the effect of several ImageNet pre-trained Resnet variant models with different network layers used as feature extractor in DeepLab model to solve semantic image segmentation task. In this study, three Resnet34, Resnet50, and Resnet101 models as network extractor of DeepLabV3 were explored. The experiment found that semantic image segmentation model performance measured by the best accuracy and average accuracies of DeepLabV3-Resnet34, DeepLabV3-Resnet50, and DeepLabV3-Resnet101 are (0.87, 0.86) (0.86, 0.84), and (0.92, 0.88) respectively. Based on the experiment, DeepLabV3-Resnet101 achieved the best semantic segmentation performance than the other models**

Keywords—semantic segmentation, Deeplab, satellite image

I. INTRODUCTION (*HEADING 1*)

Semantic image segmentation task is an exciting computer vision problem aims to label each pixel of an image with a given set of labels such as "road", "river", "tree", "building", "car". The task is known as a dense prediction as the label is predicted for every pixel in the image. Despite having similarity with instance image segmentation, semantic image segmentation is different from the former as it does not separate instances of the same class. Hence, the generated segmentation map by semantic image segmentation does not distinguish two separate objects with similar pixel category. According to [1], a plethora of semantic image segmentation methods can be categorized broadly into three types. The first type is DCNN-based systems consisting of a bottom-up image segmentation cascade, followed by DCNN-based region classification.

The second type is the segmentation model which used features extracted from DCNN model. Thirdly, DCNN model as pixel classifier without preceded by the segmentation process.

Based on their study, [1] concluded that the main challenges of semantic image segmentation using Deep Convolutional Neural Nets (DCNN) models are: (1) reduced feature resolution, (2) existence of objects at multiple scales, and (3) reduced localization accuracy due to DCNN invariance. To address these problems, [1] proposed DeepLab model as semantic image segmentation model. The author has compared the performance of a combined DeepLab model with (1) Conditional Random Field (CRF) model to refine the segmentation result (DeepLab-CRF model), and (2) Resnet101 model as a feature extractor using several image dataset. The DeepLab model is tested using PASCAL VOC 2012 and Cityscapes dataset. Performance of the model tested using the former dataset achieved mIOU 79.7; whilst, using the later dataset performed mIOU 70.4, which are the highest performance from the other tested models.

Behind its high performance as a semantic image segmentation model, DeepLab model relies on ImageNet pre-trained Deep Residual Network (Resnet for short) model [2] which serves as a feature extraction module. The feature map generated by Resnet as a Deep Convolutional Neural Nets model is used as input for DeepLab model to detect object candidate in the image segment. It is hypothesized that Resnet is an important factor that contributes to DeepLab performance. Therefore, this study aims to compare the effect of several pre-trained

Resnet models to the performance of DeepLab as a semantic image segmentation model for satellite image.

II. RELATED WORKS

A. Semantic Image Segmentation

Semantic image segmentation is formulated as a classification problem the aims to predict pixel category of an image. According to [3], the semantic image segmentation problem can be formulated as follows. Given an image as a pixel set $S = \{(x_i, y_i), i = 1,2,..,N\}$ where: x_i is a pixel value, N is the number of pixels, $y_i \in \{C_1,..,C_m\}$ is a pixel label, and m is the number of pixel class.

Given $\boldsymbol{\theta}$ be a vector of semantic segmentation model. The objective of model training is to estimate model paramaters by minimizing an objective function $\mathcal{L}(\boldsymbol{\theta}, x_i, y_i)$ using representation learning approach. Hence, the estimated parameter $\boldsymbol{\theta}$ can be represented as a solution to:

$$arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x_i, y_i). \quad (1)$$

B. DeepLab Model

DeepLabV3 model, which is the 3rd variant of DeepLab model, proposed by [1] has improved three areas of semantic image segmentation models as follows. *First*, to address the problem of diminishing spatial resolution in the feature maps generated by a sequence of max-pooling and striding at consecutive layers of Deep Convolutional Neural Nets. In contrast to previous semantic segmentation models which use deconvolutional layers, in addressing spatial resolution reduction in feature maps, DeepLab proposed to use atrous convolution. As a simple illustration of atrous convolution using one-dimensional (1-D) signals is as follows. Given 1-D signals $x[i]$ and filter $w[k]$ of length K, the atrous convolution produces output $y[i]$ defined as:

$$y[i] = \sum_{k=1}^{K} x[i + r.k]\, w[k], \quad (2)$$

where: r be a stride for sampling the input signal. Using atrous convolution is an image classification network is converted into dense feature extractors without requiring learning any extra parameters. This approach causes a DCNN training process becomes faster than the convolutional-deconvolutional approach. Second, to achieve a robust object segmentation at multiple scales by proposing atrous spatial pyramid pooling (ASPP). ASPP works by examining an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views. By using this method, the model can capture objects and image context at multiple scales. Third, to achieve a robust object localization boundary by combining the final DCNN layer with a fully connected Conditional

Random Field (CRF). The initial DeepLab model architecture proposed by [1] is then improved further by [4] [5] [6] [7].

C. Resnet Model

A study by [2] [8] concluded that adding more layers to DCNN model does not always achieve lower training and testing errors. Some evidence showed that the best ImageNet models using DCNN architecture typically contain between 16 and 30 layers. This phenomena has been studied intensively by [9] who argued that a deep structured neural network model tends to suffer from vanishing/exploding gradient. Hence, there are some potentials that deeper network models fail to perform better than their shallow network models. To address this problem, [2] proposed residual block as a new neural network layer.

The Deep Residual Nets (Resnet) model is initially proposed by [2] to build a deep structure neural network for image classification. Resnet is a very popular DCNN model that won the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015 for classification task. The successfull achievement of Resnets model is then followed by other deep learning models as a framework to ease training of the respective model.

The model consist of many stacked residual block in which each unit can be formulated in general as:

$$y_l = h(x_l) + \mathcal{F}(x_l, \mathcal{W}_l), \quad (3)$$

$$x_{l+1} = f(y_l), \quad (4)$$

where: x_l and x_{l+1} are input and output l-th unit, $\mathcal{F}$ is a residual function, $\mathcal{W}_l = \{\mathcal{W}_{l,k}: 1 \le k \le K\}$ is a set of weights and biases associated with the l-th Residual Unit, and K is the number of layers in a Residual Unit, and f is the operation after element-wise addition. The experiment by [8] suggested ReLU is used for $\mathcal{F}$ as well as f functions, and $h(x_l) = x_l$ (identity function) serves as skip connection.

$\mathcal{W}_l$ term can be implemented with 1×1 convolutions if $\mathcal{F}$ function and x_l have a different dimensionality. Should f function be an identity mapping, $f(y_l) = y_l$, then Eq. (4) becomes $x_{l+1} = y_l$ and Eq. (3) can be simplified into:

$$x_{l+1} = x_l + \sum_{i=l}^{L-1} \mathcal{F}(x_i, \mathcal{W}_i) \quad (5)$$

for any deeper unit L and any shallower unit l. The study by [8] argues that the skip connection between layers that adds the outputs from previous layers to the outputs of stacked layers makes it possible to train more deeper DCNN model than what was previously possible. Similarly [10] concluded that shortcut connections in Resnet architecture significantly lowers the difficulty of training so that performance enhancements in both model training and generalization error can be achieved. The initial Resnet model architecture proposed by [2] is then improved further by [11] [12] [13].

III. RESEARCH METHOD

A. Dataset and Data Preprocessing

The source of this experiment's dataset is Indonesia Geospatial Information Agency (BIG) that has provided SPOT 6 image and ground truth of the Pangkuh area, Kapuas District, Central Kalimantan Province, Indonesia (see Figure I). The training dataset in this study comprises 6,792 ground truth patches of 256×256×3 pixels. Data augmentation using 900,1800 and 2700 rotation is applied to the input dataset to increase labelled data. In this study, the final training dataset comprises 27,170 ground truth patches of 256×256×3 pixels.

Based on land cover, each patch is labelled as paddy field and non-paddy field. The total input dataset is divided randomly into a training dataset (80%) and validation dataset (20%). The input data, ground truth data, and segmentation output from DeepLab using several Resnet models are shown in Figure II.

B. Model Training and Testing

This study explored the performance of DeepLabV3 (DeepLab for short) as a semantic segmentation model using three ImageNet

pre-trained Resnet model, namely: Resnet34, Resnet50, and Resnet101. The DeepLab model's parameters are predicted using Adam optimization algorithm with weight decay to optimize the flatten loss of cross-entropy loss function. To avoid overfitting during model training and improve the generalization of the model, early stopping is adopted. Model performance in this study is measured using the best accuracy and average training accuracy

IV. RESEARCH RESULT AND DISCUSSION

The Semantic Image Segmentation model can be summarized in Table I, and its segmentation outputs are showed in Figure II. As Table I indicates, the accuracy of the DeepLabV3 model achieved the best accuracy compared to DeepLabV3-Resnet34 and DeepLabV3-Resnet50 models.

Table I. Training Accuracy of Each Model

Performance Metric	DeepLabV3-Resnet34	DeepLabV3-Resnet50	DeepLabV3-Resnet101
Best Accuracy	0.87	0.86	0.92
Average Accuracy	0.86	0.84	0.88

These results confirm the claim reported by [8] that the skip connection between layers improves the model generalization ability to learn from the training dataset despite its deep structure.

FIGURE I. LOCATION MAP OF THE STUDY.

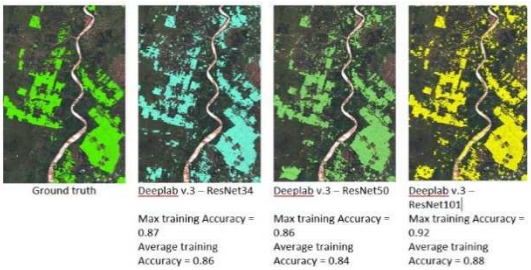

FIGURE II. COMPARISON OF SEMANTIC SEGMENTATION OUTPUTS

V. CONCLUSION

High-performance semantic image segmentation model is imperative to Land Use/Land Cover classification from satellite imagery in many application domains. For example, in agriculture, semantic segmentation is useful for generating a paddy field map from a satellite image. Resnet model becomes a critical factor in the performance of DeepLab model in semantic image segmentation. Learning from previous publications which predict deep structures of DCNN-based model will reduce model performance due to the vanishing/exploding gradient and overfitting.

However, this experiment showed some evidence that it is not always the case. With the increase in the neural network layer of Resnet model, which serves as a feature extractor network in DeepLabV3 model, more parameters need to be predicted. From an experiment using Resnet34, Resnet50, and Resnet101 combined Conditional Random Field with DeepLabV3 found that the best accuracy and average accuracy of each model is: (0.87, 0.86) (0.86, 0.84), and (0.92, 0.88) respectively. However, the experiment results raise more research questions for future research on maintaining or even improve high performance of DeepLabV3 model for semantic image segmentation for multiclass classification (labels) problem.

ACKNOWLEDGMENT *(Heading 5)*

The authors would like to thank Dr Alexander Gunawan S.G. as the Head of Mathematics and Statistics Department, Binus University and Professor Kiyota Hashimoto from Interdisciplinary Graduate School or Earth System Science and Andaman Natural Disaster Management (ESSAND), Prince of Songkla University, Phuket Campus, Thailand for providing High Performance Computing facilities. The authors also would like to thank Indonesia Geospatial Information Agency (BIG) for providing SPOT 6 image and ground truth of the Pangkuh area, Kapuas District, Central Kalimantan Province as input dataset. ESRI Indonesia supports this study under Cooperation Agreement No: 066/Dean.SCS/IV/2017 and No: 003/ESRI-EDU/PKS/2017; and funded partially by Binus University under International Research Grant No: No.026/VR.RTT/IV/2020.

REFERENCES

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[2] 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.

[3] A. Valada, G. Oliveira, T. Brox, and W. Burgard, "Towards robust semantic segmentation using deep fusion," in *Robotics: Science and Systems (RSS 2016) Workshop, Are the Sceptics Right? Limits and Potentials of Deep Learning in Robotics*, 2016, vol. 114.

[4] C. Liu *et al.*, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 82–92.

[5] B. Cheng *et al.*, "Panoptic-deeplab," *arXiv Prepr. arXiv1910.04751*, 2019.

[6] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "Deeplab-based spatial feature extraction for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 251–255, 2018.

[7] Y. Lin, D. Xu, N. Wang, Z. Shi, and Q. Chen, "Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model," *Remote Sens.*, vol. 12, no. 18, p. 2985, 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016, pp. 630–645.

[9] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 473–479, 1996.

[10] S. Li, J. Jiao, Y. Han, and T. Weissman, "Demystifying resnet," *arXiv Prepr. arXiv1611.01186*, 2016.

[11] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv Prepr. arXiv1603.08029*, 2016.

[12] T. Akiba, S. Suzuki, and K. Fukuda, "Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes," *arXiv Prepr. arXiv1711.04325*, 2017.

[13] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, 2018.

978-0-7381-4432-0/20/$31.00 ©2020 IEEE